

THE MARY E. BIVINS LIBRARY
of
THE SOUTHWEST CHRISTIAN SEMINARY
Phoenix, Arizona

**RIVERSIDE TEXTBOOKS
IN EDUCATION**

EDITED BY ELLWOOD P. CUBBERLEY

PROFESSOR OF EDUCATION

LELAND STANFORD JUNIOR UNIVERSITY

**DIVISION OF SECONDARY EDUCATION
UNDER THE EDITORIAL DIRECTION
OF ALEXANDER INGLIS**

PROFESSOR OF EDUCATION

HARVARD UNIVERSITY

RIVERSIDE TEXTBOOKS IN EDUCATION

BY THE SAME AUTHOR

**MONROE, DeVoss, AND KELLY. Educational Tests
and Measurements.**

309 pages with charts.

This is a clear and simple statement as to the nature of the different tests which have been evolved, their use, their reliability, what are the best standard scores so far arrived at, and in particular how to diagnose results and apply remedial instruction.

MONROE. Measuring the Results of Teaching.

297 pages with charts and tables.

This book presents the essential information needed by grade teachers to enable them to use standardized tests to measure and determine for themselves the effectiveness of their own instruction.

AN INTRODUCTION TO THE THEORY OF EDUCATIONAL MEASUREMENTS

BY

WALTER SCOTT MONROE

*Professor of Education and Director of the Bureau of
Educational Research, University of Illinois*



THE CINCINNATI BIBLE SEMINARY
LIBRARY

HOUGHTON MIFFLIN COMPANY

BOSTON NEW YORK CHICAGO SAN FRANCISCO

The Riverside Press Cambridge

011.26
M 7537

COPYRIGHT, 1923


BY WALTER SCOTT MONROE

ALL RIGHTS RESERVED

The Riverside Press
CAMBRIDGE • MASSACHUSETTS
PRINTED IN THE U.S.A.

TO
MY WIFE

1996



Digitized by the Internet Archive
in 2023 with funding from
Kahle/Austin Foundation

EDITOR'S INTRODUCTION

WHOLLY within the past two decades, and largely since 1909, a vast change in educational procedure has been effected in this country as the result of the development of scientific methods as applied to the measurement of both the children to be taught and the instruction to be given. It is no exaggeration to say that the introduction of these new working tools has done more than any preceding addition to educational theory or practice to transform education from an art into a science, with a well-organized body of scientific procedure to be mastered in preparation for the work. So fundamental has been the change wrought by the introduction of these new tools that school supervision is fast being changed from guesswork into scientific accuracy, and the apprentice method of preparing for the work is giving way to professional and technical preparation in schools organized for that special purpose.

Where formerly we had only a technique of classroom practice, organized in the form of a rather minute methodology, to enable us to estimate the quality and the effectiveness of classroom work, we now have tests and scales and norms which are rapidly being standardized and made accurate. Where formerly we had only the teacher's guess as to the children who were bright and average and dull, we now have graded and standardized group and individual intelligence scales for use in determining the native ability the pupils possess for mastering the work of the school. Similar scales are now being developed for use in guiding the older pupils in the choice of a life-career. Where previously we graded and promoted pupils largely on the basis

of their chronological age, we to-day can determine their mental age with accuracy and know that it is the more important basis of the two to use. Where formerly we expected pupils to master the course of study somewhat equally, and graded and advanced pupils rather evenly, we to-day understand how to calculate the achievement quotient for each pupil in the school and know that it is the real measure of ability to make school progress. Where formerly we used only simple arithmetic in tabulating results, and spoke in terms of the simple arithmetic average, we to-day have a well-organized body of mathematical knowledge and statistical procedure to apply to the analysis and interpretation of the results, and these results we express in terms of median, mode, frequency-distribution, percentiles, probable error, and degree of correlation.

So rapidly have these new procedures been evolved, due to the large number of vigorous young workers who have been attracted to this new phase of educational activity, that probably nothing to-day serves so well to differentiate the newer scientifically trained type of school supervisor from the older successful-practitioner type of educational worker as does the ability to use intelligence scales, to measure and diagnose by means of standardized educational tests, and to apply the new statistical methods to the analysis of the results obtained. The mastery of these new tools has to-day become almost essential in the equipment of any man or woman who now aspires to educational leadership as a school principal or a supervisor of instruction.

So rapidly, too, has this new knowledge been developed, organized, put into textbooks, and disseminated, that we have now reached the point where an advanced textbook on tests and measurements, dealing with the fundamental theory lying back of the construction, use, and interpretation of educational tests, is needed for the use of students

who have first become familiar with them through the means of elementary textbooks or schoolroom practice. This the present volume in this series of textbooks, written by one who has done much to develop and standardize and apply educational tests, aims to supply. It ought to prove very useful as a textbook in a second or advanced course in Educational Tests and Measurements in Teachers' Colleges and Schools of Education, and it can be read with much profit by superintendents and principals of schools who have previously become somewhat familiar with tests and testing procedure.

ELLWOOD P. CUBBERLEY

PREFACE

As we have advanced in the construction and use of standardized objective tests it has become clear that we should endeavor to refine our measuring instruments and our use of them. In order to accomplish this it is necessary that we become conscious of the assumptions upon which our measurements are based, and that we come to understand more fully the principles of the construction and use of educational tests. We need to become familiar with the technique for making critical studies of educational tests, and for evaluating them with reference to their functions.

This volume is the outgrowth of an attempt by the author to give, to advanced college and graduate students, a course which would equip them so that they would be able to make critical studies of educational tests and to form intelligent judgments with reference to their usefulness. Since the manuscript was begun, several years ago, it has undergone numerous revisions. As there was no satisfactory text available, several of the chapters have been typewritten and made available for intensive study by the members of the author's classes. During the school year of 1921-22 the first eight chapters were studied and criticized by a class of eight graduate students.

Although the volume is the direct outgrowth of the author's attempts to instruct college students, he hopes that it will afford a means of giving to superintendents and others who direct the use of educational tests in our public schools a point of view, as well as acquainting them with certain fundamental principles which seem necessary in order to secure the maximum returns from the investments that are

being made in this field. The number of available educational tests has multiplied until it is necessary for one who is formulating a testing program to make a selection of the tests to be used. This selection should be made intelligently. After the tests are selected they should be used to increase the efficiency of the school. It is the author's hope that this volume will prove helpful in assisting those working with educational tests to accomplish these ends.

Two chapters of statistical methods are included. In one sense they are not an integral part of the volume. They are intended to be used as a reference by those readers who are not sufficiently acquainted with statistical methods to understand the procedures described in connection with the construction and evaluation of educational tests. For a more complete treatment of statistical methods the reader should consult authoritative texts in this field.

An author is naturally indebted to many persons. In the present case the author feels that his greatest indebtedness is to his students, particularly those enrolled in his class in "Theory of Educational Measurements" during the first semester of the school year of 1921-22. He is also indebted to his colleague, Professor E. H. Cameron, who has read the manuscript of several of the chapters. He desires to acknowledge his indebtedness to Dean F. J. Kelly and Professor J. C. DeVoss for their permission to utilize certain materials from *Educational Tests and Measurements*, of which they are co-authors. They also have read certain portions of the present manuscript. The writer is indebted to a large number of other workers in this field. The number is too great to mention here, and appropriate acknowledgments are made in the body of the text.

WALTER SCOTT MONROE

UNIVERSITY OF ILLINOIS

November 1, 1922

CONTENTS.

CHAPTER I. THE BEGINNINGS OF STANDARDIZED OBJECTIVE TESTS 1

Origin of standard objective tests — Groups of workers contributing to early development — J. M. Rice, pioneer — E. L. Thorndike — C. W. Stone — S. A. Courtis — B. R. Buckingham — Binet — Group intelligence tests — Other events — Studies of the accuracy of school marks — School surveys — Research organizations.

Questions and topics for investigation — Selected references.

CHAPTER II. NATURE AND PROCESS OF EDUCATIONAL MEASUREMENTS 15

Two kinds of description — Scales used for quantitative description — Standard units — Qualitative characteristics described in quantitative terms — Description in terms of subjective scales — Measurement of mental abilities in terms of performance — The complete description of a performance — The process of mental measurement — Tests and scales — Units used in quantitative description of abilities are not standard — Assumptions implied in the measurement of mental ability — These assumptions only approximately true — A performance which can be observed is required — Arguments against the measurement of mental abilities — Limitations of ordinary measurements — Subjectivity of the process a source of error — Subjectivity of examination marks — Subjectivity of final "grades" — Subjectivity of ordinary measures a source of error — Other limitations of ordinary measures of abilities of pupils.

Questions and topics for investigation — Selected references.

CHAPTER III. USES OF EDUCATIONAL MEASUREMENTS IN THE WORK OF THE SCHOOL 39

General intelligence *vs.* specific abilities — Types of educational measurements — Plan of analysis to determine uses of educational measurements — School activities requiring measurement.

1. *Administrative and supervisory activities.*

Promotion and classification — Educational and vocational guidance — Evaluation of school efficiency — Rating of teachers — Reports to patrons.

2. *Instructional activities.*

Diagnosis of pupils with reference to achievements — Opportunity for diagnosis varies with the grades — Varies also with content of instruction — Diagnosis with respect to study procedures — Need for diagnostic study tests.

3. *Research activities.*

Determination of procedure through measurement of outcomes — Summary.

Questions and topics for investigation — Selected references.

CHAPTER IV. THE CONSTRUCTION OF EDUCATIONAL TESTS 56

General structure of educational tests — Types of exercises — Devising tests to give objective record — Definition of difficulty — Types of tests produced — Power tests and rate tests — Requirements governing construction — Nature of ability — The development of ability — Relation of performance to ability.

1. *Requirements for constant functional relationship.*

How secured — Excluding other abilities — Controlling other *x*-factors.

2. *Requirements for unrestricted functioning.*

Type of ability to be measured — Limitations of functioning due to content — Limitations due to structure of the test — Uniform, scaled, spiral, irregular, and cycle performances — Exercise must not be ambiguous.

3. *Agreement with educational objectives.*

Should measure effectiveness of instruction — Exception in case of a school survey — Prior determination of objectives — Relation of scaled performance test to educational objectives.

4. *Requirements for the administration of a test and description of the performance.*

Elements in a complete performance.

5. *Control of testing conditions.*

Attitude of examiner — Preparation by pupils — Distribution of test papers — General data — Explanation of test to pupils — Timing the performance — Illustrative test directions.

6. *Description in terms of significant dimensions.*

Law of the single variable.

7. *Equal spacing of exercises on the scale of difficulty.*

Selection of exercises for the test — Content — Suitableness — Difficulty — Assumptions underlying construction of scaled tests — Selection of exercises for a scaled test — Relation between per cent of correct responses and difficulty — Finding the inter-grade interval — Exercise method — Quartile method —

Distribution method — Combining the three methods — Locating the zero point — Determining the final scale values.

Questions and topics for discussion — Selected references.

CHAPTER V. DESCRIPTION OF THE PERFORMANCES OF PUPILS 106

Quantitative scale required — Rate of uniform performances — Rate of non-uniform performances — Two measures of quality — Accuracy of uniform performances — Scale of accuracy — Standards of accuracy — Arbitrary standards — Two types of uniform performances — Equivalent standards of accuracy necessary for comparable measures — Accuracy of non-uniform performances — Weighting exercises — Scores based on unweighted performances — Description of a scaled performance — Relation of accuracy to difficulty — Standards of accuracy — Estimating the level of difficulty corresponding to a given standard — Kelley's shorter method of estimating scores — Van Wagenen's shorter method — Describing a scaled performance in number done correctly — Relation between number done correctly and difficulty score — Sum of difficulty values used as scores — Combined scores — Accuracy when number of answers is limited.

Questions and topics for investigation — Selected references.

CHAPTER VI. DESCRIPTION OF PERFORMANCES OF PUPILS: QUALITY SCALES 133

Description when quality does not mean accuracy — Constructing a quality scale — Collecting and selecting samples — Quantitative description of quality of samples — Modifications of the method — Location of zero point — Subjectivity of these methods — Objective methods of scale construction — Using a quality scale.

Questions and topics for investigation — Selected references.

CHAPTER VII. DERIVED SCORES 145

Lack of a common unit of measurement — Lack of a common zero point — Need for — Derived scores in terms of common units and point — Variability of chronological age-groups as a common unit — A proposed common zero point — McCall's derived scores — Pintner's derived scores — Derived measures expressed in terms of grade norms — Derived measures in terms of grade percentile scores — Measures expressed as achievement ages — Quotients as derived measures — Significance of the achievement quotient.

Questions and topics for investigation — Selected references.

CHAPTER VIII. NORMS FOR EDUCATIONAL TESTS AND THEIR USES 161

Definition of a norm — Types of norms — General limitation of norms — The basis of satisfactory norms — Norms stated with reference to testing conditions — Effect of acquaintance with a test — Effect of coaching — Equivalence of duplicate forms — Errors of interpretation *vs.* errors of measurement — Different uses require different norms — Use of norms illustrated; two variables — Three variables — Norm in more complex situations — The type of norms required for evaluating the efficiency of a school — General intelligence *vs.* school grade and achievement — Pupil diagnosis — Grade norms for different mental ages — Mental-age norms superior to grade norms — Comparison with mental-age norms by ratio or achievement quotient — Advantages of mental-age norms.

Questions and topics for investigation.

CHAPTER IX. HOW TO MAKE A CRITICAL STUDY OF AN EDUCATIONAL TEST 182

General outline for such a study — Title — Nature of pupil's performance — Description of pupil's performance — Function of the test.

Validity of the test.

Validity of physical measurements — Determination of validity complex — Variability of performance in spelling and arithmetic — Variability in accuracy shown by a table — Functional relation of performance to ability — Functional relation between score and performance — Between score and ability — Methods of determining validity.

Objectivity in describing performances.

The "personal equation" in testing — Constant and variable errors — Errors involved in scoring reproductions.

Reliability.

Obtained and true scores — Methods of determining reliability — Coefficient of reliability — Relation of coefficient of reliability to length of test — Calculation of coefficient of reliability from one application of test — The index of reliability — Probable error of measurement — Calculation of P.E.M. of a test; Formula A — Application of method — Reducing one set of scores to scale of another — Calculating P.E.M. of a test; Formula B — Probable error of measurement expressed as a ratio — Relation of P.E.M. to length of test — P.E. of class scores — Reliability in terms of coefficient of correspondence — Overlapping of successive grade groups.

Discrimination.

Certain criteria for — Objections to use of large units — Comparison with criterion measures — Teachers' marks — The validity of prognostic tests — Measures yielded by other tests — Comparison with composite test scores — Inferences concerning validity — When a test lacks in validity — Validity of significance — Norms — Practical considerations.

Questions and topics for investigation. Selected references.

CHAPTER X. USING STANDARDIZED EDUCATIONAL TESTS . 232

Standardized tests not teaching devices — Uses in teaching.

1. Selection for specific purposes.

Definition of need — Validity — Norms — Cost of testing material — Time cost of testing and scoring.

2. Administration.

Testing by a single teacher — Administering a testing program in a school system — Scoring the papers — Training in the use of a quality scale — Scoring by pupils or clerks.

3. Planning remedial procedure.

Interpretation of scores — Types of errors — Graphic representation of results — Remedial procedure — Use of tests by teacher and supervisor.

4. Diagnosis of pupil difficulties with tests.

Diagnosis with reference to grade norms — Diagnosis in terms of achievement quotients — Significance of achievement quotient — General diagnosis *vs.* detailed diagnosis — Analytical diagnosis.

5. Useful remedial measures.

Planning remedial instruction — Pupil diagnosis with reference to study procedure — Setting immediate educational objectives — Motivating school work by use of tests — Reporting the achievements of pupils to parents.

Uses of standardized educational tests by the supervisor.

Six major supervisory uses for tests — Promotion and classification — Placement of new pupils — Promotion of pupils — Classification within a grade — Selection of exceptional children — Educational and vocational guidance — Objective measuring in the supervision of instruction — Evaluating the efficiency of a school system — Intelligence measurement and efficiency evaluation — Use of comparative data — The measurement of progress — Modifying factors — The rating of teachers — School publicity — Scientific experimentation.

Questions and topics for investigation.

CHAPTER XI. THE IMPROVEMENT OF EXAMINATIONS . 277

Examinations not completely replaced by tests — Making examinations more objective — Increasing the objectivity of the marking by definite rules. — By using questions which permit of only one correct answer — True-false exercises — Yes and No exercises — Recognition exercises — Completion exercises — Advantages of the "new examination" — Limitations of "new examination" — Unequal difficulty of questions not a serious defect — Agreement with minimum essentials — Recognition of significant dimensions — Norms for ordinary examinations subjective — "Grades" *vs.* measures of achievement — Translating achievement quotients into school marks — Translating point scores into school marks — Objective norms for examinations.

Questions and topics for investigation.

CHAPTER XII. STATISTICAL METHODS: THE FREQUENCY DISTRIBUTION AND ITS DESCRIPTION 296

Purposes of statistical methods — Knowledge of significance of derived facts — Sources of error in — Plan of treatment to be followed — The frequency distribution — Continuous and discrete series — Method of expressing facts in a continuous series — Assumptions concerning frequency distributions — Questions arising in their formation — Shape to be expected — Description of a frequency distribution — The central tendency — The average — Calculation by short method — Calculation when intervals are irregular — The median — Median in discrete and in continuous measures — Calculation from a frequency distribution — Calculation of special cases — Calculation of percentile points in a distribution — Mode — Measurement of variability of a frequency distribution — Calculation of average deviation — Standard deviation — Other measures of variability — The median deviation, or P.E. — Relationship between measures of variability — Interpretation of measures of variability — Effect of chance errors on the central tendency — Effect of using a sample on the central tendency — Comparison of two averages — Exercises.

CHAPTER XIII. STATISTICAL METHODS: RELATIONSHIP EXISTING BETWEEN SETS OF PAIRED FACTS 336

Sources of paired facts — Constructing a correlation table — Calculation of coefficient of correlation by the Pearsonian formula — The four steps — Interpretation of — Effect of variable and constant errors — Effect of sampling — Computing P.E. of coefficient of correlation, due to sampling — What a relationship

CONTENTS

xix

between measures means — The departure from perfect correlation — The regression equation — The probable error of estimate — The error of estimate represented graphically — Effect of selection of data on — Probable error of measurement — Coefficients of probable errors of estimate and probable errors of measurement — Exercises.

INDEX 363

LIST OF TABLES

1. Relation between per cent of correct responses and difficulty of exercises	66
2. Kelley's shorter method of calculating difficulty scores applied to Trabue's Language-Completion Test	124
3. Record sheet for recording judgments of judges concerning the quality of samples	136
4. Form of recording the rankings of samples by judges in order to secure per cent of "better judgments"	139
5. Differences in quality corresponding to per cent of "better judgments"	140
6. Per cent of pupils who maintained the same accuracy of performance through two groups of examples	192
7. Teachers' ratings of compositions using Nassau County Supplement to Hillegas Scale	197
8. Subjectivity of scoring reproductions by the word-counting method	199
9. Correlation table showing relation between the number of examples attempted on Form 1 of the Addition Test, Series B, and Form 2 of the same test	204
10. Per cent of pupils failing in one large building of a city school system	260
11. Distribution of deviations from a "standard" mark of two sets of teachers' marks on fifth-grade arithmetic papers, by two methods of marking	281
12. Silent reading rate scores of 81 seventh-grade pupils as determined by Starch's Reading Test No. 6	299
13. Scores given in Table 12 grouped in intervals to form a frequency distribution	300
14. Frequency distribution of facts in Table 13, grouped according to a different scale of intervals	307
15. Illustrating calculation of average by short method	310
16. Showing calculation of average for one case of irregular intervals	312
17. Illustrating calculation of the median of the frequency distribution	315

18. Illustrating calculation of the median when the total of the frequencies is an odd number	317
19. Illustrating calculation of median when it falls at junction point of two intervals	319
20. Rate scores on Starch's Silent Reading Test	320
21. Per cent of accuracy; Curtis Arithmetic Test, B, Division	321
22. Illustrating the calculation of the average deviation of a frequency distribution	323
23. Illustrating the calculation of the standard deviation of a frequency distribution	326
24. Correlation table for rate scores in silent reading	338
25. Table illustrating perfect correlation	342
26. Showing actual and estimated scores on Curtis Silent Reading Test No. 2	350
27. Coefficients of probable errors of estimate and probable errors of measurement	355
28. Effect of nature of trait and selection of population group upon coefficient of correlation	358

LIST OF FIGURES

1. Distribution of measures of silent reading ability as measured by the Kansas Silent Reading Tests	32
2. Distribution of marks of teachers of English and History in University High School, Chicago	33
3. Graphical representation of a normal distribution, showing relation between number of pupils and abilities	94
4. The Ogive Curve, showing relation between accuracy and difficulty	120
5. Relation between point scores and achievement quotients in comprehension of silent reading	247
6. Gains in comprehension of silent reading, due to definite instruction for this purpose	254
7. Gains in rate of silent reading due to definite instruction for the purpose	255
8. Achievement ages for arithmetic and for reading for two cities	266
9. Median mental age and median I.Q. for each grade for the same two cities	267
10. Achievement quotients for each grade for the same two cities	268
11. Distribution of pupils according to mental age	306
12. Calculation of the coefficient of correlation between scores derived from Form 1 and Form 2 of the Illinois General Intelligence Scale	339
13. Correlation of Form 1 scores with Form 2 scores of the Illinois General Intelligence Scale, fifth grade	353

AN INTRODUCTION TO THE THEORY OF EDUCATIONAL MEASUREMENTS



CHAPTER I

THE BEGINNINGS OF STANDARDIZED OBJECTIVE TESTS

Origin of standardized objective tests. Although the achievements of pupils have been measured in some form since schools were organized, the construction of standardized objective tests for this purpose is a comparatively recent development. These measuring instruments have two distinguishing characteristics. *First*, they tend to be objective, which means that the measures which they yield are relatively independent of the person who makes them. In other words, different examiners will obtain approximately the same results when measuring the abilities of the same pupils under the same conditions. *Second*, these tests are standardized, so that norms are available for interpreting the measures which they yield. For example, if the measure of a given achievement of a fifth-grade pupil is 41, this magnitude can be compared with the norm or standard for the fifth grade. By doing so we can ascertain whether this pupil is above standard, just up to standard, or below standard with respect to this particular achievement.

The idea of standardized objective tests is relatively simple. It is likely that the idea had been comprehended, in

part, by many teachers before it was developed by recent workers. Professor E. L. Thorndike¹ has called attention to a plan for measuring the achievements of pupils which was used by an English schoolmaster as early as 1864. The instrument, called a "scale book," was crude, but it included the germ of many of the ideas which are incorporated in our present-day educational tests. Sample performances, representing various degrees of excellence, were collected and arranged in the form of a scale similar to our present handwriting scales. Such "scale books" were constructed for several school subjects, and they appear to have been used in the school where the idea originated. However, the work of this English schoolmaster, Reverend George Fisher, appears to be isolated from the development of educational tests in this country. In this connection, it should be noted that the construction of many of our present-day measuring instruments was not possible until appropriate statistical methods had been worked out. In this field Sir Francis Galton was the pioneer. The beginning of standardized objective tests in this country dates from the work of Dr. J. M. Rice, in 1894-95.

The measurement of general intelligence, or capacity to learn, was developed separately from the measurement of the achievements of pupils. Many attempts were made to "measure the mind" by indirect methods. Phrenology represented one attempt which attracted much attention. Physiognomy was another. Still later, certain physical measurements were proposed by some physiologists as a means of learning about a child's mind. Closely connected with the physical measurements were those of reaction time and other mental functions. It, however, remained for

¹ See *Journal of Educational Psychology*, vol. 4, pp. 551. A quotation from *The Museum, a Quarterly Magazine of Education, Literature, and Science*, vol. III, 1864, is reproduced as a communication.

Binet, in 1905, to propose the method which is now generally used. His work marks the beginning of our present measurement of general intelligence. Group intelligence tests are a very recent development.

Two groups of workers contributed to the early development. The early development of standardized objective tests for measuring the achievements of pupils was contributed to by workers in two fields, psychology and school administration. In studying the functions of the human mind, the psychologists frequently devised tests in school subjects. These tests were more or less crude, but, generally, they were objective. Standards were seldom determined, since they were of little value to the psychologists. The school administrators were interested in questions of courses of study, school organization, methods of instruction, etc. Until a quarter of a century ago, such questions were dealt with on a controversial basis. The side which had the best "debaters" won. More recently, educators have attempted to answer these questions in a scientific manner and, in doing this, objective tests have frequently been devised.

The development of standardized objective tests is so recent, and so many workers have contributed to the movement, that it is difficult to select the persons and the events which deserve recognition. It is an interesting fact that all persons who have made major contributions are living at the time this is written (1922). However, a few persons and events appear to be distinctive. A number of these are mentioned below.

J. M. Rice, the pioneer, 1894-95. In America the first person whose work has been given a place in the development of standardized objective tests is Dr. J. M. Rice. His tests were not standardized, or at least he made no use of them as such. Since the test papers were rated under his immediate supervision, and in one investigation the test was

given by him in person, the objectivity of the tests was not an issue. Rice's contribution to educational measurement was, rather, in the use of tests in certain school subjects for answering questions concerning the best course of study, time allotment, methods of instruction, and the like. In doing this he focused attention upon the idea of measuring the achievements of school children.

Rice's first investigation (1894-95) was in the field of spelling. His first test consisted of a list of fifty words. This was sent to a number of school superintendents, of whom twenty responded, sending in the returns from more than sixteen thousand school children. He constructed a second test, in sentence form. This test was given under Rice's personal supervision, and the scoring was checked by his assistants. Among the conclusions which he drew from the results, Rice stated that pupils who had spent forty minutes a day for eight years in studying spelling did not appear to possess any greater ability to spell than those who had devoted only ten minutes a day to the subject. When he reported his investigation, the educators of that day, with few exceptions, vigorously attacked the implied thesis that it is possible to measure the results of teaching spelling by ascertaining how well the pupils can spell certain words. This was done in spite of the fact that teachers had been and still were attempting to measure the results of their instruction in spelling by means of written tests and by other crude devices. A principle that was generally accepted in practice was rejected when utilized as a means of answering a disputed question concerning the optimal time to be devoted to the teaching of spelling. It was claimed that the object of teaching spelling "was not to teach pupils to spell, but to develop their minds." Later Rice gave tests in arithmetic and language.

E. L. Thorndike, 1903, 1904, 1909-10. Professor E. L. Thorndike has made many notable contributions to the de-

velopment of standardized objective tests. He also has contributed indirectly to the movement through the work of his students. At the time of the first of Rice's work, Thorndike was a student of statistical methods. He was also deeply interested in the field of psychology. This combination of interests began to bear fruit as early as 1903. In this year, in collaboration with Fox, he published a study entitled *Relation between the Different Abilities in the Study of Arithmetic*.¹ In studying this relation, a number of tests on the operations of arithmetic were used. The motivation of this contribution was psychological rather than administrative. In this respect, it differs from the work of Rice, whose motive was largely that of the school administrator.

In 1904 Thorndike published the first edition of his *Mental and Social Measurements*. In addition to an account of statistical procedure, this volume contained many of the principles upon which the construction of our present tests is based. It was revised in 1913, but the revision consists, primarily, of adding concrete illustrations of the principles. This book is yet an important source of information for workers in this field, and for a number of years was essentially the only source.

In December, 1909, Professor Thorndike presented his handwriting scale before the meeting of Section L of the American Association for the Advancement of Science. The report of the derivation of this scale was published in the *Teachers College Record* the following March. The appearance of this scale was an important event in the development of educational tests. This scale is a device which makes possible the description of the general merit of handwriting in quantitative terms. This characteristic of a pupil's handwriting is essentially qualitative but, by means of Thorndike's Handwriting Scale, it can be described in quan-

¹ *Columbia University Contributions to Philosophy, Psychology, and Education*, pp. 32-40 (February, 1903).

titative terms. By the use of the theorem that differences in quality which are equally often noted by competent observers are equal, he secured a set of specimens of handwriting which differed from each other by known amounts of quality. These specimens were arranged in order of increasing merit to form the scale. The quality of a sample of a pupil's handwriting is measured by matching it with the division of the scale which it most nearly resembles in general merit. The appearance of this scale is significant also because it was constructed without special reference to a study which called for the use of a measuring instrument. Although Thorndike was not indifferent to the usefulness of his handwriting scale, he was primarily interested in the technique of its construction.

The principles underlying the construction of Thorndike's Handwriting Scale were applied to the field of English composition by Hillegas, in 1912. Since that time they have been used with some modification and elaboration by a number of other workers.¹

C. W. Stone, 1908. The work of Rice inspired C. W. Stone, then a student under Thorndike, to undertake the study of two questions in the field of arithmetic:

- (1) What is the nature of the product of the first six years of arithmetic work?
- (2) What is the relation between distinctive procedures in arithmetic work and resulting abilities?

In the work of Stone we have a combination of the interests of the psychologist and of the school administrator. The former is due to Stone's contact as a student with Thorndike, and the latter to the inspiration received from the work of Rice.

In the study of the above questions Stone devised two tests in the field of arithmetic, one on the fundamentals, and

¹ See Chapter VI for an account of the method of scale construction.

the other on reasoning. These tests were objective, in that printed directions were provided for giving them and for the scoring of the test papers. These directions were sufficiently detailed to insure that different workers would obtain approximately the same results in applying the tests to the same pupils. The tests, as used by Stone, were not standardized. Since the completion of the original study the reasoning test has been standardized, and used extensively in school surveys and by teachers. In the construction of his tests, Stone was more scientific than Rice. Weights were determined for examples which obviously differed in difficulty, so that the units of the scale would be more nearly equal.

S. A. Courtis, 1909. S. A. Courtis cooperated with Stone by giving his arithmetic tests. Stone's study was confined to the sixth grade. Courtis conceived the idea of giving the tests in all grades, including the high school. This he did during the school year of 1907-08. He states that he was interested in measuring the growth of pupils in arithmetic, and in establishing norms of attainment for the several grades. He found the Stone tests unsatisfactory for this purpose, and during the following year he devised a group of arithmetic tests which he called Series A. These were made available for use in September, 1909. The series consisted of eight tests. There was one test on the fundamental combinations or tables in each operation. The fifth test called for the copying of figures, and Test No. 7 included examples from the four operations. Two tests were devoted to reasoning or the solving of problems. A great deal of care was exercised in the construction of the tests, and in the formulation of the directions and record blanks for their administration. A time limit was set, so that measures of the rate of work were secured. The scoring was made objective, and the tests were standardized.

The purpose of Courtis in devising this series of tests was

different from that of earlier workers in the field. His interests were neither those of the psychologist nor those of the school administrator but, rather, those of the instructor. He thought of the tests as instruments to aid the teacher in instructing pupils. They were to do this by measuring progress of the pupils, and the established norms were to be used by the teacher as objectives. The comparison of the measures of achievement with the norms furnished a diagnosis of the pupils with respect to their educational needs.

The extensive use of these tests in schools convinced Courtis that they were unsatisfactory in a number of respects. For this reason he devised and made available, during the year 1913-14, a new group of tests called Series B. It is this group of tests that is generally referred to as the "Courtis Standard Research Tests in Arithmetic." The series consists of four tests, one on each of the fundamental operations with integers. The examples are relatively "long" and within each test all are the same size. In the addition test each example consists of three columns of figures of nine figures each. The tests measure both rate and accuracy.

L. P. Ayres, 1912, 1915. L. P. Ayres contributed to the development of educational measurements by the construction of his handwriting scale (three-slant edition) in 1912, and his spelling scale in 1915. He has, also, contributed to the movement through numerous addresses. The Ayres Handwriting Scale (three-slant edition) represents an ingenious attempt to secure an objective index of the quality of handwriting. Legibility was adopted as the criterion of quality. Samples of pupils' handwriting were read, under controlled conditions, and an index of legibility was calculated from the average rate of reading. This method of scale construction has not been followed by other makers of handwriting scales because investigation revealed that Ayres's scale is not distinctly superior to Thorndike's, and the

method of construction used by the latter is much simpler. Ayres's most significant contribution is his spelling scale. In it he emphasized the importance of determining the educational objectives as a prerequisite of test construction.

B. R. Buckingham, 1913. In 1913 B. R. Buckingham, a student under Thorndike, published the account of the derivation of his spelling scale. This work is largely theoretical and the scale has not been widely used. It is, however, significant historically because it represents a new type of measuring instrument. Words were selected and their difficulty was determined on the basis of the percent of correct spellings by pupils in the various school grades. These words were then arranged in the order of their difficulty. This procedure produced a measuring instrument which began with words which practically all pupils could spell, and increased in difficulty until, at the other end of the scale, relatively few pupils were able to spell the words correctly. The idea underlying this type of measuring instrument is that the ability of a pupil may be measured in terms of the level of difficulty which he is able to reach on the scale.

This principle of test construction has been followed by a number of workers in this field, notably, Woody in the operations of arithmetic, Hotz in algebra, Henmon and Brown in Latin, Trabue in language, and Van Wagenen in history.

Binet (1905), 1908. All of the events up to this point have been in connection with the development of instruments for measuring the achievements of pupils in school subjects. Because of the intimate relation of general intelligence to achievement, and because, as we shall show later, there is need for combining measures of intelligence and achievement, it is appropriate to mention briefly the development of general intelligence tests.

In 1904 Binet, an eminent French psychologist, was appointed by the French Minister of Public Instruction, as a

member of a commission for the organization of classes for subnormal children. This brought Binet face to face with the problem of devising some means for determining what children were subnormal. In the endeavor to solve this problem he devised, in 1905, in collaboration with Simon, the group of tests which have become known as the "Binet General Intelligence Tests." They were revised by Binet in 1908. Binet's most significant contribution was the use of age-norms in interpreting the performances of pupils in terms of mental age.

The work of Binet attracted the attention of American psychologists, and his tests have been revised and improved by a number of our psychologists, notably, Goddard and Terman. The Stanford Revision by Terman is, perhaps, the most widely known and used.

Group intelligence tests (1918). The Binet tests must be given to pupils individually. Since frequently more than an hour is required to administer them, it is obvious that their use is necessarily limited. Prior to the entry of the United States into the World War, in the spring of 1917, a number of psychologists had been experimenting with tests that might be given to pupils in groups. A. S. Otis, a student under Terman, had completed the derivation of such a test. When it was decided to introduce psychological testing into the United States Army, he generously placed his results at the disposal of the committee which had this work in charge. The derivation and extensive use of group intelligence tests for psychological examining in the United States Army is probably the most important event in connection with this type of test. Since the cessation of hostilities there have appeared a very large number of group intelligence tests, which are of the same general type as that originated by Otis and adopted for use in the United States Army.

Other events. Many other events might be mentioned. In fact, the number of recent events is so great that one is unable to evaluate them properly. A great deal of ingenuity has been exhibited by the makers of silent reading tests. Notable among these are E. L. Thorndike, who has contributed the Thorndike Scale Alpha II for measuring the understanding of sentences, and the Thorndike Visual Vocabulary Test; and F. J. Kelly, who devised the Kansas Silent Reading Test. The latter event is particularly important because of the simplicity of the test and its consequent widespread use. A very recent development, which at the present time appears to be of major importance, is the introduction of derived scores, and the comparison of measures of achievement with measures of general intelligence so as to obtain achievement quotients. These two topics will be discussed in detail in later chapters.

Studies of the accuracy of the school marks. The development of educational tests has been stimulated by a number of events not having to do with their construction. Notable among these are studies of the accuracy of school marks.¹ These investigations revealed that school marks were highly subjective, and hence inaccurate. Perhaps the most important event in this connection was the publication, in 1914, of a monograph by F. J. Kelly, entitled, *Teachers' Marks*.² In this monograph the author summarized the existing studies of school marks, and added the results of a number of new investigations. This accumulation of evidence left no doubt that teachers' marks were inaccurate, and the demonstration of this fact created the need for instruments which would yield more accurate measures of achievements.

School surveys. The survey movement has sustained a twofold relation to the development of educational tests.

¹ Certain of these studies will be described in Chapter II.

² Teachers College Contributions to Education, No. 66.

12 THEORY OF EDUCATIONAL MEASUREMENTS

School surveys stimulated derivation and use of educational tests. In turn, the present type of school survey has been made possible by the derivation of instruments for measuring the abilities of pupils. The study of a school system by a committee appointed for that purpose is not a new activity. One writer mentions the appointment of an educational commission of the City of Chicago, in 1894, as one of the earliest of these endeavors. The use of the word "Survey" in connection with the study of a school system appears to date from a "Survey" of Pittsburgh, in 1907. Following this date school surveys shortly became common events. The inquiry into the public schools of New York City, in 1911-12, is notable because it is the first survey in which educational tests were used to measure the achievement of pupils as a means of evaluating the efficiency of a school system. S. A. Courtis was appointed as a member of the survey commission, and gave Series A of his arithmetic tests to over thirty thousand children. Since this date there have been few school surveys in which educational tests have not been used. In a number of school surveys, those making the study of the school system felt that the existing tests were unsatisfactory, and devised new instruments to meet their needs. Notable illustrations of this are the Cleveland Survey Arithmetic Tests, the Willing Composition Scale, and the Nassau County Supplement to the Hillegas Scale.

Research organizations. The development of educational measurements has been greatly facilitated by the establishment of research organizations. These have been of two types. One has been organized in school systems, primarily for the purpose of directing and stimulating the application of educational tests. Sometimes the organization has been evolved out of the activities carried on by some division of the superintendent's office. In other cases it has represented the creation of an essentially new department placed, gen-

erally, under the control and direction of the superintendent. One of the first of these organizations was in New Orleans, in 1912, but the department has since been discontinued. The same year a department was established in Rochester, New York. In 1913 a Bureau of Research and Reference was established in New York City. In the following year similar organizations were established in Kansas City, Missouri, and Boston, Massachusetts.

The other type of research organization has been established by state educational institutions or state departments of public instruction. The Bureau of Measurement and Efficiency, at the University of Oklahoma, was established in 1913. This appears to have been the first of such organizations. A number of others were established very soon after. The Bureau of Coöperative Research at Indiana University, the Bureau of Educational Measurements and Standards at Kansas State Normal School, Educational Service, Extension Division at the State University of Iowa, and the Bureau of Educational Measurements at the University of Nebraska were established in 1914. The State Department of Public Instruction of Wisconsin established a similar organization in 1915. More recently, organizations of this general type have been established in a large number of other state educational institutions, and recognition has been given to the work by the department of public instruction in a number of states.

This type of research organization has been very influential in popularizing the use of educational tests. Through them, school superintendents and teachers have been made acquainted with the tests, and frequently they have been trained in the use of the tests. Service has also been rendered by making the tests easily accessible, and by standardizing tests devised by other workers. In many instances those directing research organizations have devised new tests.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What were the motives that actuated the pioneers in educational measurements?
2. What reasons can you give for the attack that was made upon Rice's work?
3. How do you explain the fact that the test movement has been carried on almost entirely by young men?
4. Trace the development of standardized objective tests for the different school subjects.
5. How does Ayers's Handwriting Scale differ from the one devised by Thorndike?
6. What contribution has Courtis made in this field?

SELECTED REFERENCES

- ASHBAUGH, E. J. "Organization and Function of a Bureau of Educational Research"; in *School and Society*, vol. ix, pp. 577-84 (May 17, 1919).
- AYRES, L. P. "History and Present Status of Educational Measurements"; in *Seventeenth Yearbook of the National Society for the Study of Education*, part II, chap. I.
- AYRES, L. P. "Measuring Educational Progress through Educational Results"; in *School Review*, vol. xx, pp. 300-09 (May, 1912).
- BALLARD, P. B. *Mental Tests*, chap. I. Hodder and Stoughton, Ltd., London, 1920.
- BUCHNER, E. F. "School Surveys"; in *Report of the U.S. Commissioner of Education for 1914*, vol. I, chap. XXIV.
- BUCKINGHAM, B. R. *Spelling Ability: Its Measurement and Distribution*. Teachers College Contributions to Education, No. 59.
- COURTIS, S. A. "Measurement of Growth in Efficiency in Arithmetic"; in *Elementary School Teacher*, vol. x, pp. 58-74, 177-99 (October and December, 1909); vol. XI, pp. 171-85, 360-70 (December, 1910, and March, 1911).
- HOLMES, HENRY W. "A Descriptive Bibliography of Measurement in Elementary Subjects"; in *Harvard Bulletins in Education*, vol. v, June, 1917.
- JOHNSTON, J. H. "A Brief Tabular History of the Movement towards Standardization by Means of Scales and Tests of Educational Achievements in the Elementary School Subjects"; in *Educational Administration and Supervision*, vol. II, pp. 483-492 (October, 1916).
- RICE, J. M. *Scientific Management in Education*. New York, 1913.
- STONE, C. W. *Arithmetical Abilities and Some Factors Determining Them*. Teachers College Contributions to Education, No. 19, 1908.
- THORNDIKE, E. L. "The Measurement of Educational Products"; in *School Review*, vol. xx, pp. 289-99 (May, 1912).
- THORNDIKE, E. L. "Handwriting"; in *Teachers College Record*, vol. II, March, 1910.
- WYLIE, A. T. "A Brief History of Mental Tests"; in *Teachers College Record*, vol. XXIII, pp. 19-33 (January, 1912).

CHAPTER II

NATURE AND PROCESS OF EDUCATIONAL MEASUREMENTS

Two kinds of description. We apply two kinds of description to physical objects. When a thing is described in terms of its color, texture, form, beauty, and so forth, we have *qualitative* description. When we tell how large a thing is, the description is *quantitative*. Take, for example, a table. The kind of wood of which it is made, the color of its finish, and its shape are terms which may be used in its qualitative description. A statement of its weight, height, dimensions of its top, and other information which tells how large it is describes it quantitatively.

To measure a thing is simply to describe it in quantitative terms. Whenever we determine how much or the amount of any characteristic of an object, we have measured that object with respect to this characteristic.

Scales used for quantitative description. All quantitative description requires a scale of quantity.¹ The most common type of scale is one which is formed by laying off the unit of measurement from a starting-point. This starting-point is called the zero point, and usually means not any of the characteristic which is named in the unit. This scale may be extended indefinitely to meet the needs of measurement. An object is measured or described by determining how far it extends along this quantitative scale from the zero point. Thus, when the height of a tree is described as being twenty-three feet, this means that the height extends along the scale

¹ By a liberal interpretation "quantitative description" may be taken to include description in terms of large, very large, heavy, tall, etc. It is used in a more restricted sense in this book.

of linear measure a distance of twenty-three units (feet) from the zero point on the ground.

It is necessary that the unit which is used in forming the scale be of the same kind of magnitude as that which we are attempting to measure. For example, we cannot use the scale of which the square foot is the unit to measure the weight of an object. Furthermore, the magnitude which we are attempting to measure must be consistent throughout. These requirements are so obvious in the case of physical measurements that it may appear unnecessary to mention them. Attention is called to them here because similar requirements must be satisfied in the measurement of mental traits.

Another type of scale which may be used in quantitative description is one consisting of a series of objects of a given type arranged in order of increasing magnitude. The scale objects may be numbered in order, starting with the one of least magnitude, or they may be given other convenient numerical designations. When this has been done, the numerical name of a scale object defines a definite magnitude. Quantitative description by means of such a scale consists in estimating which of the scale objects the object being described most nearly resembles in magnitude. The position of the scale object in the scale describes the other quantitatively. If the scale objects differ from each other by known amounts of the magnitude being considered the quantitative descriptions obtained by using the scale will have a more exact meaning. The measurements yielded by such a scale will be more easily handled and interpreted if the differences between scale objects are equal. If, in addition, the distance of the smallest magnitude from an absolute zero point is known, measurements by means of such a scale may become as exact as those obtained by using a scale formed by laying off units from a zero point.

A quantitative scale of this type could be devised for measuring the height of boys by selecting a group of boys of appropriate heights and arranging them in the order of increasing height. If they were selected so that the differences in height between adjacent boys in the scale and the zero point were known, the height of any other boy could be described quantitatively by determining which scale-boy he most closely resembles in height. The boys of the scale could be numbered 1, 2, 3, 4, 5, etc. The height of a boy not included in the scale would be described quantitatively when it is stated that his height is equivalent to that of boy No. 11 in the scale.

Standard units. The magnitude of the units which we commonly use in the quantitative description of physical objects has been fixed by custom, or otherwise, and for this reason we speak of them as standard units. For example, the magnitude of the yard has been fixed in the United States by federal law. It is the distance between two points on a metal bar which is kept at Washington. The fixing of standard units is not essential to the process of measurement. Any magnitude may be arbitrarily defined as a unit by anyone. An instrument constructed accurately on the basis of such a unit would yield as accurate measurements as those which we secure now from instruments based on our standard units. Such measures would, however, be understood only by those who were acquainted with the unit used. It is, therefore, obviously desirable that, in the measurement of physical objects, standard units be used.

Qualitative characteristics described in quantitative terms. Under certain conditions, characteristics of objects which are essentially qualitative may be described in quantitative terms. The quality of a sample of handwriting is a qualitative characteristic. In order to describe this characteristic in quantitative terms it is only necessary to

construct a scale consisting of a series of suitable samples of handwriting, arranged in order of increasing quality. The scale positions of the samples in this array can be defined numerically by assigning a suitable number to each sample.¹ The quality of any given sample of handwriting could then be described in quantitative terms by determining the scale-sample which it most nearly resembles in quality, and using the numerical name of this scale position as the quantitative description of the quality of the sample of handwriting.

This type of scale is useful for certain purposes; but, as we shall show in a later chapter, it is difficult to construct such a scale, and it is even more difficult to match objects which we desire to measure with the correct scale objects. It should, however, be recognized that we have here a device which, theoretically, makes possible the quantitative description of all characteristics of objects, both quantitative and qualitative.

Description in terms of subjective scales. Another method of expressing the description of a qualitative characteristic in quantitative terms is to estimate the degree to which the quality of the characteristic in question approaches perfection. For example, the quality of a sample of handwriting may be described as 73 per cent of perfection. Here the description is in terms of a scale similar in certain respects to the one described above. The significant difference is that such a scale exists in the mind of the person who is making the description and has no objective manifestation. Such a scale may be described as subjective. Different persons are likely to set up different scales. Consequently, they will give different descriptions of the same thing.

Measurement of mental abilities in terms of performance.

¹ The procedure to be followed in constructing this type of scale is described in Chapter VI.

We can know of mental ability only as it is manifested in some performance. Ability is ability to do. If it cannot function, if it cannot produce action, it is not ability in the sense in which the term is here used. Thus, it is appropriate that we describe quantitatively or measure ability in terms of the performance which its functioning produces under specified conditions.

It has been contended that we are not able to measure mental ability, that we can measure only performance. This distinction, however, is not significant for practical purposes. As observed in the above paragraph, the meaning which we attach to mental ability comes from the performance which it produces. When the conditions¹ under which the ability functions are specified and controlled, i.e., are the same for all children, the measurement of ability and the measurement of performance have synonymous meanings. If these limitations are kept in mind we may speak interchangeably of measuring ability and of measuring performance. For every performance there is a corresponding degree of ability.

The complete description of a performance. A pupil's performance is completely described when three things are stated concerning it:

1. The amount of the performance or the rate at which it was given.
2. Quality or accuracy of the performance.
3. The character of the exercise in response to which the performance was given. In certain cases this description may be given the difficulty of the exercise or exercises.

These three dimensions are implied in the questions: "How fast can a pupil work?" "How well can he do the tasks set?" "How hard tasks can he do?" The relation of a

¹ "Conditions" here include mental states and the procedure which the children use in applying their abilities.

performance to the ability of which it is the manifestation is not defined until the conditions under which the performance was given are specified. Hence, a statement of the conditions under which the ability functioned is an essential part of its measurement. The three characteristics or dimensions of a performance are not always equally important, and some measuring instruments are constructed so that they yield a separate description of only one or two of the characteristics. Other instruments provide for combining the descriptions of the rate of performance with the quality, and some yield scores which are a combination of all three dimensions.

The process of mental measurement. Briefly, the measurement of an ability consists of securing a quantitative description of the performance which that ability produces under specified and controlled conditions. Since this ability has produced this performance, it is assumed that, under the same conditions, it would produce an equivalent performance. From this point of view we may think of the measurement of ability as consisting of securing the quantitative description of the performance which it will produce under given conditions; but the description of this predicted performance is, at the same time, a description of a performance that has been given.

Tests and scales. Both of the words, "test" and "scale," have been used in naming our instruments for the measuring of mental abilities. The measurement of an ability involves two distinct steps: first, securing from the pupil a performance; and, second, describing that performance. If we wish to use the words "test" and "scale" in a precise way, we should think of a test as being that portion of a measuring instrument which is used to secure the performance from the pupil. The function of a scale is realized in the description of the performance. The complete measuring instru-

ment is, strictly speaking, a test plus a scale. In naming our measuring instruments these words have been used without much discrimination. In so far as there has been any discrimination with respect to the precise meanings of the words, the term has been used which is characteristic of the distinguishing feature of the measuring instrument. For example, we have the Courtis Arithmetic Tests, the Kansas Silent Reading Tests, and the Thorndike Handwriting Scale.

Units used in quantitative description of abilities are not standard. For describing performances of pupils there are no standard units. A large number of units have been used. As a rule, the makers of our measuring instruments have used the units which were most convenient. In describing the rate of reading the word or line is generally used as the unit. In the operations of arithmetic, the example is the usual unit. Words differ in respect to length and to meaning; arithmetical examples differ in length and complexity. Hence, these units are not standardized in the sense that the yard is a standard unit for linear measure. One example may not be equal to another example, but a yard is always the same. It would be very helpful if a group of standard units could be adopted for the measurement of mental abilities; but this has not been done. Certain proposals of standard units have been made, and these will be described in a later chapter on the description of pupils' performances.

Assumptions implied in the measurement of mental abilities. As indicated above, the procedure of measuring mental abilities and the interpretation of the resulting measures imply certain assumptions concerning the nature of mental abilities and their functioning. Some of these assumptions have already been suggested, but it will be worth while to give definite statement to the significant assumptions in this place. Some of these assumptions depend upon the character of the measuring instruments; some of them

are introduced by reason of group measurements; others depend upon the use of common norms for the interpretation of all measures yielded by a given measuring instrument.

1. *It is assumed that the performance sustains a constant functional relation to the ability which is being measured.* This assumption is implied in the use of common norms for the interpretation of all measures yielded by a test. For example, if the norm of a given test for the fourth grade is 17, the measures of the achievements of all fourth-grade pupils are interpreted by comparison with this norm. This is done even though the achievements of the pupils are measured by different examiners at different places and at different times. This assumption means that any change in ability from pupil to pupil produces a proportional change in the performances, and that all variations in performances are produced by corresponding changes in ability.

2. *It is assumed that the performance is the product of the functioning of certain mental processes or abilities and no others, at least not in a way to disturb the constant functional relation.* This assumption is implied in the one just mentioned above. In case the performance is to sustain a constant functional relation to a given ability it is obvious that this performance must not be affected by other abilities, unless the contributions of these other abilities are synchronized with the fluctuations produced by the ability being measured. Generally speaking, different abilities are sufficiently independent so that their contributions do not correlate perfectly. This being the case, it follows that the performance secured must be the product of the functioning of certain mental processes or abilities. If other abilities contribute, their contributions must be slight unless the constant functional relation is to be grossly disturbed.

3. *It is assumed that we can secure the functioning of a given ability for the purpose of measurement.* This assumption is

implied in our use of standardized objective tests. When a test is given to a group of pupils we assume that we have secured the functioning of the same ability in the case of each pupil. We, furthermore, assume that each pupil has had a reasonable opportunity to demonstrate his ability, unless it is obvious that he has not. In other words, our customary administration of standardized objective tests implies that we can secure the functioning of a given ability whenever we desire it. It is only necessary to present the test to the pupils and we shall secure the functioning of the ability that we desire to measure.

4. *It is assumed that the testing conditions can be controlled, not only through a single testing period, but also on subsequent occasions and by different examiners.* This assumption is fully included in the first one, but it appears to be sufficiently important to justify a repetition in this form. Testing conditions include the explanation of the test, the timing of it, the attitude of the pupils, the effort which they make, the lighting of the room, and even the temperature and atmospheric pressure. In fact, testing conditions include all factors which may affect the performances of pupils. It is obvious that if these factors are not controlled, so that their effect upon the performance is the same for all pupils, even when they are tested at different times and by different persons, the constant functional relationship between the performance and the ability being measured will be disturbed.

5. *In the measurement of general intelligence it is assumed that all pupils tested have had equal opportunities to acquire the abilities for which the test calls.* Although the function of a general intelligence test is to yield measures of the native or inherited abilities of pupils, it does this indirectly by measuring certain acquired abilities. Most of the exercises that are included in general intelligence tests make demands upon school training and upon the general experience of pupils.

This assumption should be interpreted in terms of the average of all the abilities for which the test calls. It is obvious that, in the case of the abilities called for by a particular sub-test or a particular exercise, all pupils may not have had equal opportunities to acquire them. If the entire test is considered as a whole this assumption is approximately in agreement with actual conditions for children living in the same general environment.

These assumptions only approximately true. These assumptions are only approximately true. A perfect constant functional relation between the ability being measured and the performance that is secured is never realized. In some cases this relation is disturbed by the effect of other abilities, even in cases where it may appear that the supplementary abilities have been reduced to a minimum. For example, it has been demonstrated that the ability to write figures may seriously affect the performance of pupils upon certain types of arithmetic tests.¹ The functional relation is also disturbed because we are not able to exercise complete control over testing conditions. This is especially true when the test is administered by different persons and at different times.

The fact that these assumptions are not fully realized introduces errors in the measures we obtain and in their interpretation. Gradually, our measuring instruments are being improved so that these assumptions are more closely approximated. The important thing is to recognize what assumptions are implied in our procedure, and to know the extent to which failure to realize completely these assumptions affects the measures which we secure. In a later chapter, under the head of validity, we shall consider in detail the procedure

¹ See Thorndike, E. L., and Courtis, S. A. "Correction Formula for Educational Testing"; in *Teachers College Record*, vol. xxi, pp. 1-24 (January, 1920).

for ascertaining the extent to which the assumptions are realized in the case of a particular test.

A performance which can be observed is required. It is, of course, obvious that, for the purposes of measurement, it is necessary to secure a performance which can be accurately observed. Preferably, it should be one which culminates in a permanent record. Unfortunately, the functioning of some mental abilities does not normally result in such performances. For example, the functioning of the ability to read silently does not result in a performance that can be observed. In order to secure an observable performance in the measurement of this ability it is necessary to introduce additional abilities. This violates the second assumption stated above, and thus tends to disturb the constancy of the functional relation between the ability and the performance. When this occurs the resulting measurements are likely to be lacking in validity.

Arguments against the measurement of mental abilities. Largely because of the assumptions and difficulties involved, some have contended that the measurement of mental abilities is not possible. In support of their position they have asked whether an idea has weight, length, breadth, or height. When asked in this form only a negative answer is possible. Hence, these critics say an idea cannot be measured. The influence of a good teacher, they have claimed, cannot be measured. They contend that a mother's love cannot be described in quantitative terms. One fault of those who have subscribed to this position has been that they were thinking of the measurement of mental abilities in terms of the most difficult and indefinite cases with which we have to deal. They have, also, failed to recognize in their thinking the fact that we define ability in terms of the action which it produces. Hence, when we measure a performance we are indirectly measuring the ability which func-

tioned in producing it. Furthermore, they have failed to realize that since the beginning of the human race, mental traits of its members have been measured. We are constantly measuring our friends and our associates with respect to certain traits. A teacher is constantly measuring the abilities of his pupils. Such measurements are frequently crude, and not always in satisfactory terms. It may be that we shall never realize our ideals of validity and accuracy in the measurement of certain mental traits. It appears that human activity, and particularly the carrying on of the work of our schools, require the measurement of various human traits. We have always measured them, and it is absurd to contend that the measurement of mental abilities is impossible. As Professor Thorndike well states, "Whatever exists at all exists in some amount." Measurement is simply the description of this amount.

The limitations of ordinary measurements of mental ability. As was pointed out in Chapter I, the measurement of mental abilities is not new. The work of our schools requires that abilities of pupils be measured. However, the measurements which we have been making are subject to significant limitations. Since educational tests are proposed as instruments for securing more satisfactory measurements, it is appropriate that we inquire into the limitations of the ordinary measurements which teachers are accustomed to make.

Subjectivity of the measuring process is a source of error. The process of measurement is said to be *subjective* when the result obtained depends in a large way upon the person making the measurement. This is the case when different persons measuring the same thing obtain different results. When different persons measuring the same thing obtain approximately the same result the resulting measures are described as *objective*. Perfect objectivity is not obtained,

even in the refined measurement of physical objects. It is much less completely realized in the measurement of mental abilities even when our most refined tests are used. Hence, the term "objective," when applied to a test, must be interpreted to mean that the test is relatively objective when compared with other instruments which are subjective in a high degree.

If two or more different measurements are obtained for the same thing, only one result can possibly be right. It may be that all are incorrect. Hence, we shall discuss the accuracy of our present school marks in terms of subjectivity and objectivity. If we find that different teachers, employing similar procedures in measuring the same ability, obtain different results, we must conclude that their method of measurement is subjective. To say that a certain method of measurement is subjective is to say that it tends to yield erroneous measures. Some of the measures may not involve significant errors, but if the process is highly subjective most of them will involve errors that cannot be neglected. Other causes may contribute to the total error of the measures, but proof of their subjectivity is sufficient.

The subjectivity of examination marks. The written examination is one of the most common means of measuring the abilities of pupils. It requires that the pupil exercise certain abilities in producing a written record of his responses to questions and other exercises which the teacher or some other school official has prepared. The quantitative description of this written record is the pupil's examination "grade" or mark. It may be in terms of percent or of some other symbol, such as "A," "B," etc.

Without considering the question of whether a pupil's performance on an examination is a true indication of his ability, we may inquire concerning the degree of subjectivity of the marks which teachers assign to the examination papers.

When the pupil's performance is a written record we may have it described, i.e., measured independently by a number of teachers. If their descriptions ("grades") agree closely, we must conclude that this portion of the process of measurement by means of the written examination is highly objective. On the other hand, if the "grades" assigned to the same paper by different teachers differ widely, we must conclude that these "grades" are highly subjective and, in consequence, cannot be accurate measures of the abilities of the pupils.

The subjectivity of examination marks has been studied by having an examination paper duplicated and then asking competent teachers, working independently, to "grade" it. Perhaps the best known of these investigations are those by Starch and Elliot,¹ of the University of Wisconsin. They studied the marking of examination papers in three school subjects, English, geometry, and history. Their method and the nature of the facts revealed may be illustrated by an account of their study of the marking of a geometry paper.

A facsimile reproduction was made of an actual examination paper in plane geometry. A copy of this reproduction was sent to each of the high schools included in the North Central Association of Colleges and Secondary Schools, with the request that it be marked on the scale of one hundred per cent by the teacher of geometry. The teacher was asked to mark the paper by the method he was accustomed to use. When we consider that the subject-matter of geometry is quite definite, and that the papers were marked by teachers who were thoroughly acquainted with the subject, it would seem that we might expect the marks or "grades"

¹ Starch, Daniel, and Elliot, E. C. "Reliability of Grading High-School Work in English"; in *School Review*, vol. xx, pp. 442-57. "Reliability of Grading Work in Mathematics"; in *School Review*, vol. xxi, pp. 254-59. "Reliability of Grading Work in History"; in *School Review*, vol. xxi, pp. 676-81.

placed upon this examination paper to be in close agreement. However, exactly the opposite was found to be the case. Returns were received from 116 teachers. Two teachers gave a "grade" above 90, while one "grade" was below 30. Twenty were 80 or above, while twenty other marks were below 60. Forty-seven teachers assigned a mark of passing or above, while sixty-nine teachers decided the paper was not worthy of a passing mark. Not only were similar results obtained by Starch and Elliot in English and in history, but their conclusions have been verified many times by other investigators.¹

In the face of such facts only one conclusion is possible; namely, that under ordinary conditions the marks assigned to examination papers by teachers are highly subjective. Such marks can represent only very crude and very inaccurate measures of the abilities of pupils. It is not too much to say that the mark which a pupil receives on an examination paper depends to a large degree upon the teacher who "grades" the paper.

It has been shown also that the same teacher is not consistent in his own marking. If a teacher "grades" a set of papers a second time, the two sets of marks will vary widely.²

The subjectivity of "final grades." Teachers compute "final grades" in a variety of ways, but they are intended to be a quantitative description of the sum total of the achievements of pupils within a given subject-matter field. Because, under normal conditions, no two teachers have equal opportunities for measuring the sum total of the abilities of a pupil in the same school subject, we cannot study the subjectivity of "final grades" by the same method that was used in studying the marking of examination papers. Two

¹ See Kelly, F. J. *Teachers' Marks*, p. 51, and following, for accounts of other investigations.

² See Starch, Daniel, *Educational Measurements*, p. 9.

methods which have been used may be illustrated by the following investigations.

1. *Kelly's investigation.* One method has been to compare "final grades" given to the same pupils by different teachers. The limitation of this method is that the two sets of measurements cannot be made at the same time. Since for a given school subject a pupil is assigned to one teacher, it is necessary to compare the "final grades" of two successive semesters or years. The most significant investigation of this type is the one made by F. J. Kelly in the public schools of Hackensack, New Jersey.¹

In this school system the pupils from four ward schools went to a common departmental school (junior high school) for the work in the seventh and eighth grades. The pupils who in the sixth-A class were taught by four different teachers were taught in the seventh-B class by the same teacher in a single school subject. A comparison of the "final grades" given to these pupils in the seventh-B class with the ones which they received in the sixth-A class will indicate the extent to which their "final grades" were subjective. If the average of the "final grades" given to the pupils in one of the ward schools (A) was higher than the average of the "final grades" given in another (B), we should expect the pupils from the first ward school to receive the higher "final grades" when they go to the common departmental school, provided the "final grades" which they received in the sixth-A class were accurate measures of their abilities. However, if the teacher in school A "graded high," and the teacher in school B "graded low," i.e., if their "grades" were subjective, the average of the "final grades" of the pupils from school A may be expected to be below the average of those from school B in the seventh-B class.

¹ Kelly, F. J. *Teachers' Marks*, Teachers College Contributions to Education, No. 66.

This latter condition is what Kelly found to exist. He states that "for work which the teacher in school C (one of the ward schools) would give a mark of 'G' in language, penmanship, or history, the teacher in school D (another ward school) would give less than a mark of 'F.'" This means that the "final grades" given by teachers are subjective, i.e., dependent, in part, upon the teacher who gives them. Another teacher would give different "grades" for the same quality of work.

2. *Johnson's investigation.* Another type of investigation of the subjectivity of "final grades" has been made by Johnson,¹ principal of the University High School of the University of Chicago. It is based upon the assumption that, when accurate measurements are made of any ability of a large representative group of pupils, the resulting measures are distributed, that is, arranged along the scale of measurement, in a certain definite way. For example, in Fig. 1, there are represented graphically four distributions of the measures of silent-reading ability secured by giving the Kansas Silent-Reading Tests. The number of measures represented in each grade is over five thousand. The base line of the curve in each case represents the scale of the test, 0, 1, 2, 3, 4, 5, and so on. At any point of this base line the height of the broken line curve above the base line represents the number of pupils having the measure represented. The general shape of these four broken line curves is the same. A few pupils received very low measures and a few very high ones. The great majority of the measures are grouped near the middle where the curve is highest. A curve which, beginning with the low measures, rises gradually and then falls gradually, as do those shown in Fig. 1, is called a "normal

¹ Johnson, F. W. "A Study of High-School Grades"; in *School Review*, vol. xix, pp. 13-24. See, also, Kelly, F. J., *Teachers' Marks*, pp. 11, and following, for reports of similar investigations.

curve." If the shape of the curve representing the distribution of a particular set of measures differs materially from

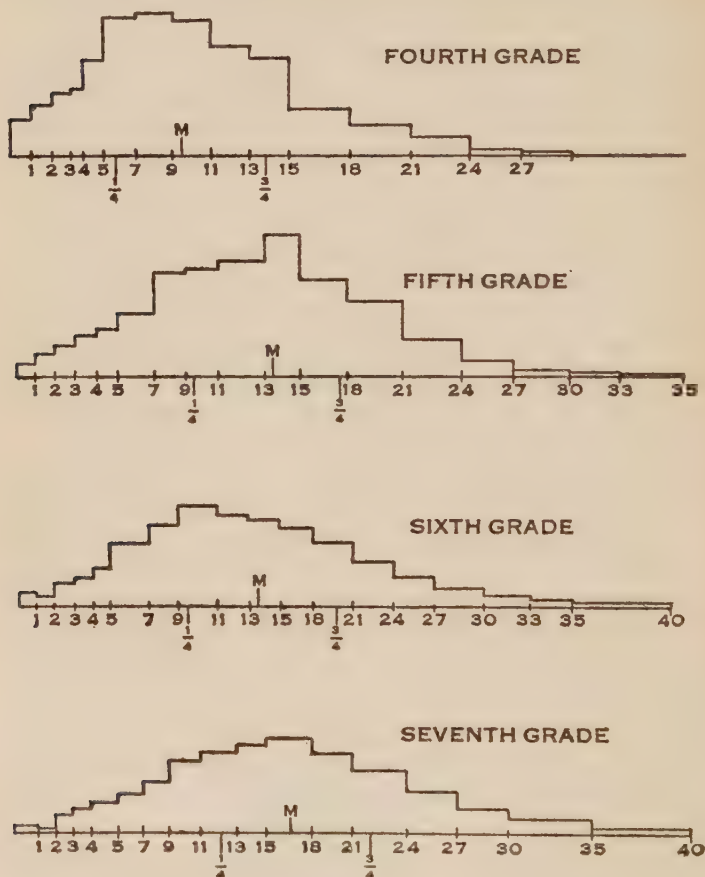


FIG. 1. SHOWING THE DISTRIBUTION OF MEASURES OF SILENT-READING ABILITY AS MEASURED BY THE KANSAS SILENT-READING TESTS.

the general shape of the curves in Fig. 1, there is reason for questioning the accuracy of the measures.

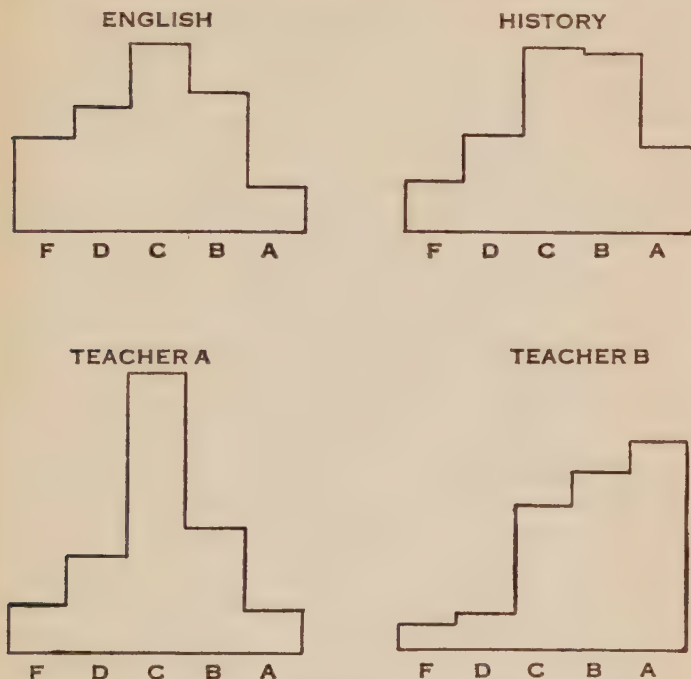


FIG. 2. (UPPER) SHOWING DISTRIBUTION OF MARKS IN UNIVERSITY OF CHICAGO HIGH SCHOOL IN ENGLISH AND HISTORY. (LOWER) SHOWING DISTRIBUTION OF MARKS OF TWO TEACHERS IN THE SAME DEPARTMENT. (AFTER JOHNSON.)

In the University High School, "F" denotes failure, and the four successive ranks above failure are indicated by "D," "C," "B," and "A." For the several departments of the school, Johnson tabulated the number of times each mark was given during the years 1907-08 and 1908-09. The con-

ditions which he found to exist may be illustrated by Fig. 2. The upper figure shows the distribution of marks in English (left) and history (right). It will be noted that in the case of English a much larger proportion of low marks ("F" and "D") were given than in history. For the high marks ("A" and "B") just the reverse is true. Both curves fail to conform closely to the normal curve described above, which suggests that the marks are dependent, in part, upon the group of teachers who gave them, i.e., are subjective.

However, the most significant part of the figure is the *lower*, which represents the distributions of the marks given by two teachers in the same department. The distribution for teacher A conforms reasonably closely to the normal curve, but that for teacher B departs from it in a very conspicuous fashion. If we assume that the two groups of pupils were equivalent with respect to their capacity to learn, the subjectivity of the marks is apparent. They depend upon the teacher who gave them. It is obvious that teacher B is accustomed to give "high grades." In so doing, he has furnished evidence that his marks are probably inaccurate.

Subjectivity of ordinary measures of the abilities of pupils a large source of error. Both "examination grades" and "final grades" have been shown to be highly subjective; that is, it has been shown that different teachers when measuring the same thing tend to secure different results. This means that the "examination grades" or other school marks assigned by any teacher are likely to be inaccurate. If they are accurate it is largely a matter of chance. There are other sources of error, but this is probably one of the largest.

Other limitations of ordinary measures of the abilities of pupils. In addition to their subjectivity, there are a number of other limitations of ordinary measures which should be noted. Some of these are not a source of error, but rather cause misleading interpretations of the measures. Such

limitations, however, are significant because measures of abilities are of value only as they are correctly interpreted.

1. *Important abilities not always chosen for measurement.* The field of a school subject includes a large number of abilities. Careful analyses are necessary in order to determine what these abilities are, and which ones are most significant. In formulating the questions of ordinary examinations the teacher does not, usually, have at hand a statement of the important abilities within the field of a school subject, and, as a consequence, some of the important abilities are frequently omitted in the measurement without being recognized as omitted. Thus, measures of a few abilities are frequently interpreted as being measures of the entire field of abilities. For example, measures of the abilities of pupils to do certain types of examples in arithmetic have been interpreted, when important types of examples were not included, as being measures covering the whole field of the operations of arithmetic. The making of satisfactory measurements is dependent upon a careful analysis of the subject-matter field in which they are being made. The ordinary classroom teacher does not have at hand these analyses in preparing examinations and in assigning school marks.

2. *No objective norms for interpreting measures.* A measure has no meaning until it is compared with a norm. The size of the measure of a pupil's ability does not at once tell whether a pupil rendered a good performance or not. A norm is something which we may use as a basis of comparison. As generally used in connection with the measurement of abilities of pupils, a norm represents the average ability of a certain group of pupils. The comparison of a measure with it thus tells one whether the pupil possesses average ability, greater than average ability, or less than average ability.

In order to be useful these norms must be stated objec-

tively; that is, they must be stated in such a way that all teachers will understand them alike. For interpreting ordinary measures there are no objective norms, and a teacher must set his own; that is, a teacher uses a norm that is subjective. This means that it is likely to be different from the corresponding norm held by another teacher. Furthermore, a teacher's norm probably does not remain constant, but changes from time to time.

In the case of an examination which is marked on the scale of 100 per cent, the norm is partly expressed in the passing mark, partly in the examination itself, and partly in the plan of marking the examination papers. The difficulty of the questions and the plan of marking them are intended to be such that a pupil who possesses the abilities covered by the examination, in a barely satisfactory degree, will receive the passing mark. Assuming that the examination yields a reasonably accurate measure of the pupil's abilities, if the list of questions is more difficult than the teacher considered it to be, the pupil will be judged to have a less degree of ability than he really has. He will be considered to be below passing when he should receive a mark above passing, simply because the examination was too difficult. This means that the teacher's norm was too high. The plan of marking the examination papers has a similar influence. A severe plan of marking will result in lower grades than a more liberal one.¹

3. *Important dimensions of abilities frequently not measured.* The complete description of a pupil's performance requires a statement of the rate at which it was produced, its quality, and the type or difficulty of the exercises in response to which

¹ It should be noted that a grade, such as "85 per cent," or "B," or "fair," is a statement of comparison of the measure of the ability of the pupil with the teacher's norm. Thus, "grades" should be defined as interpreted measures; i.e., measures compared with the norm.

it was given. These characteristics of a performance may be called its "dimensions." In ordinary measurement the rate of performance is generally neglected. In many cases, particularly in the "tool" subjects of the elementary school, the rate is important. This condition is illustrated by handwriting, silent reading, and the operations of arithmetic. To omit the measurement of the rate in such cases is to neglect an important dimension of the pupil's abilities. In some subject-matter fields the rate of work is much less important. For example, in solving problems in arithmetic the rate of work is less important than the rate at which the operations are performed. In such cases failure to measure the rate is not a serious omission.

4. *Ordinary measures are not diagnostic.* The typical school examination covers a variety of topics. Thus, the "grade" which a pupil receives is an average or general measure. It does not furnish an index of his particular weaknesses. It does not tell in what topics he is strong and in what topics he is weak. The ordinary examination is not diagnostic. For certain purposes, diagnostic measures are required. In such cases the general measures yielded by ordinary examinations have a limited usefulness.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What is measurement?
2. What do we measure by means of our educational tests?
3. What assumptions are involved in the thesis that we measure the abilities of pupils?
4. What are the characteristics of the abilities of pupils which are significant from the standpoint of their measurement?
5. What are the arguments of those who contend that we cannot measure mental abilities?
6. What are the answers to these contentions?
7. What is a test? What is a scale?
8. What is a standard unit? Are the units used in educational measurements standardized?
9. What do we mean by a standardized test?

10. What are the essential steps in the process of educational measurement?
11. What do we need to tell about a pupil's performance in order to completely describe it?
12. What types of information may we secure about a pupil's mental processes?
13. What are the limitations of teachers' marks?
14. When is a measure subjective? How would you investigate the subjectivity of teachers' marks?

SELECTED REFERENCES

AYRES, L. P. "Making Education Definite"; in *Proceedings of Second Indiana Conference of Measurement*, 1915, pp. 85-96.

BALLARD, P. B. *Mental Tests*, chap. i. Hodder and Stoughton, Ltd., London.

BROWN, WILLIAM, and THOMSON, G. H. *The Essentials of Mental Measurement*, chap. i. Cambridge University Press.

BURGESS, MAY AYRES. *The Measurement of Silent Reading*. Department of Education, Russell Sage Foundation.

COURTIS, S. A. *The Gary School. Survey Report*; volume on *Measurement of Classroom Products*, chap. viii. General Education Board.

COURTIS, S. A., and THORNDIKE, E. L. "Correction Formulæ for Addition Tests"; in *Teachers College Record*, vol. xxi, pp. 1-24 (January, 1920).

MERIAM, J. L. "Measuring School Work in Terms of Life out of School"; in *School and Society*, vol. v, pp. 339-42 (March 24, 1917).

Proceedings of National Education Association, 1918. For objections to and uses for mental tests, see pp. 298-304.

RICE, J. M. *Scientific Management in Education*, pp. 282. New York, 1913.

THORNDIKE, E. L. *Mental and Social Measurements*, chaps. i and ii.

THORNDIKE, E. L. "The Nature, Purposes and General Methods of Measurements of Educational Products"; in *Seventh Yearbook of the National Society for the Study of Education*, part ii, 1918.

CHAPTER III

USES OF EDUCATIONAL MEASUREMENTS IN THE WORK OF THE SCHOOL

General intelligence vs. specific abilities. Before considering the uses of educational measurements in the work of the school, it will be helpful to make certain distinctions relative to the types of measurements which may be made of the abilities of pupils. There is, first of all, the measurement of general intelligence. For practical school purposes general intelligence may be thought of as the measurement of the pupil's general capacity to do the work of the school. Some authorities agree with Thorndike that the result of the education provided by the school is the engendering of a number of specific or relatively unrelated abilities in the pupil. The acquisition of specific abilities implies corresponding specific capacities to learn. When the theory of specific abilities is accepted, a pupil's general intelligence is considered to be the average of selected samples of his specific capacities to learn. Other authorities contend that intelligence is essentially general, or that it includes some general factors. Neither of these concepts of general intelligence is incompatible with the working definition stated above. Most of our group intelligence tests consist of several different sub-tests. A pupil's total score is taken as the measure of his general intelligence.

If we accept the theory of specific capacities to learn we then need instruments which will yield measures of specific intelligence. Such measures are indices of the pupil's capacity to learn certain particular things. For example, the Rogers Prognostic Tests in mathematics were designed to

measure the specific capacity of pupils to learn secondary mathematics. Such an instrument may be called a specific intelligence test to distinguish it from a general intelligence test. Such tests have also been called *prognostic tests*.

Types of educational measurements. The majority of the tests which have been devised for school use have for their function the measurement of achievement, or what the pupil has learned. They are called *achievement tests*. Many achievements appear to be specific. In such a field as the operations of arithmetic we engender not one general achievement, but a large number of achievements which are relatively independent of each other. For example, a pupil may attain a very high degree of achievement in addition of integers without attaining anything like an equivalent degree of achievement in division. In fact, each of these achievements is general, in a sense, for it has been shown that a pupil may attain one degree of achievement in doing long-column addition, and exhibit a distinctly different degree of achievement in doing short-column addition. Since ability tends to be specific, we may recognize two types of achievement tests; first, those which measure separately certain specific achievements; and, second, those which yield average or general measures of a number of specific achievements within a given field. A test which yields separate measures of specific abilities is frequently called a *diagnostic test*, because it points out or diagnoses the specific weaknesses of pupils.

Instruments for measuring the abilities of pupils to engage in the various types of learning furnish us with a fifth type of test. A pupil may possess a high degree of native ability, but not be efficient in certain types of learning. For example, a pupil may have adequate capacity to learn, but be inefficient in memorizing because he goes about this work in an inefficient way. There are certain best methods of

learning, and, until the pupil has been trained in these methods, he is not likely to be efficient as a learner. Instruments yielding measures of this kind would tell us the degree to which pupils know how to engage in the various types of learning which the school requires of them. Relatively few tests have as yet been devised which have this function. Some of our silent-reading tests approximate this function for certain types of learning. For example, in the Thorndike-McCall Silent-Reading Test, the pupil is given a small assignment to read for the purpose of answering certain specific questions. This represents one type of learning. The Van Wagenen Reading Scales for history and general science furnish another illustration of this type of test. There appears to be no reason why tests cannot be devised for measuring the effectiveness of pupils in other types of learning, once the need for them is recognized.

There is some overlapping between the different types of measurements which we have enumerated. A measure of achievement is indirectly, also, a measure of intelligence, or capacity to learn. A pupil who does not achieve highly is usually one who has a relatively low capacity. In general, there is a high degree of correspondence between achievement and intelligence. This correlation is not perfect by any means, because achievement is also influenced by the effort which the pupil makes, the instruction which he has received, and by other factors. Likewise, there is a positive correlation between a pupil's achievements and the degree of effectiveness which he has attained in the methods of learning which are required. A high degree of achievement cannot be expected from a pupil unless he is efficient as a learner. There is, also, a relation between the degree of effectiveness in doing certain types of learning and the pupil's intelligence.

Plan of analysis to determine the uses of educational measurements. The primary purpose of educational meas-

urements, whether made by written examinations and other crude methods or by standardized objective tests, is to increase the effectiveness of the work of our schools. The construction of standardized objective tests and their use can be justified only if we are able to show that the more accurate measures of the abilities of pupils which they yield will increase the efficiency of our schools. Hence, the first step in determining the need for educational measurements is to inquire concerning the phases of the work of the school which require the measurement of the abilities of pupils.

School activities requiring measurement. There are a number of activities involved in the work of our schools in which measurements of the abilities of pupils are required. In some of these activities, it is necessary to have information concerning the pupil's capacity to learn; in others, we must know concerning his achievements; and, in still others, we need to know concerning the degree of his effectiveness in doing certain types of learning. The school activities which require measurement may be considered under three major heads:

1. Administrative and supervisory activities.
2. Instructional activities.
3. Research activities.

The activities which we shall consider under these headings require the measurement of the abilities of pupils if the activities are to be carried on in an efficient manner. In the absence of standardized objective tests, such as have been devised during recent years, the measurements required by these activities have been secured by teachers and other school officials by employing crude methods, such as the traditional written examination. In considering the need for standardized objective tests which shall be uniformly used by all teachers, it is necessary to bear in mind the possibility that the measures obtained by the crude methods

may be as effective in some, or even all, of these activities as the measures obtained by standardized objective tests. Thus, when we show a need for the measurement of abilities of pupils, it does not necessarily follow that we have established a need for standardized objective tests.

Some of these activities are not equally important in all grades, and all of them do not require the same type of information. We shall consider these activities in some detail with reference to the type of information which they require and also with respect to their importance in the different grades. We shall thus be able to indicate, in a general way, the types of measurements which are most important in the different divisions of our schools.

1. Administrative and supervisory activities requiring measurement

Promotion and classification of pupils for the purposes of instruction. Our public schools are usually organized with twelve grades. When a pupil enters school he is assigned to the appropriate grade. At the end of the term or year pupils who have done the work of their grade satisfactorily are promoted to the next grade. Occasionally, a pupil receives a double promotion, and, in a very few cases, a pupil is sent back to a lower grade. In some school systems there are occasional cases of promotion during the term. The promotion of pupils has been and is now based largely upon measures of their achievements. Within recent years there has been a plea to recognize, also, a pupil's general intelligence or capacity to learn. Thus, in this activity there is need for both measures of achievement and measures of general intelligence.

Within a school grade pupils may be classified in sections. Sections may be formed so that pupils possessing approximately the same capacity to learn are brought together for

purposes of instruction.¹ In some school systems special classes are formed, such as classes for gifted children, or classes for dull and backward children. It has been urged that the classification of pupils within a school grade should be based primarily upon measures of their capacity to learn or their general intelligence. However, measures of achievement should not be neglected in this connection. There are some pupils who achieve more highly than their general intelligence indicates. This is because of unusual application, or other conditions favorable to learning. There are, also, pupils of the opposite type who fail to achieve as highly as their general intelligence indicates should be expected of them. Hence measures of achievement as well as measures of intelligence are needed in the classification of pupils.

Educational and vocational guidance. This activity is closely related to the promotion and classification of pupils; in fact, below the seventh grade, educational guidance is little more than promotion and classification. In the seventh and eighth grades and, more particularly, in the high school the pupil has some choice of the subjects which he shall pursue. Thus, there is an opportunity to advise him with respect to the work which he shall undertake. Measurements of general intelligence appear to furnish the most important item of information on which to base educational guidance. The adviser also needs to have measures of the pupil's achievements as well as other information. Advice to pupils concerning the choice of a vocation should be based upon information concerning vocations, but frequently measures of pupils' general intelligence and school achievements will be valuable.

Evaluation of school efficiency. It is generally recognized

¹ The author has intentionally avoided any consideration of the desirability of classifying or segregating pupils on the basis of their general intelligence. This question will be considered briefly in a later chapter.

that a high degree of school efficiency cannot be expected unless there is an accounting or checking up at intervals. The school is, essentially, a manufacturing plant. The pupils are the raw material; the school buildings, grounds, equipment, and textbooks are the plant; the teachers and supervisors are the workmen in this educational factory. The efficiency of the enterprise depends upon the output, that is, the achievements of the pupils. It is also necessary to know the quality of the pupil-material with which the school is working. Therefore, in evaluating the efficiency of a school, we must measure the general intelligence of the pupils as well as their achievements. From one point of view, methods of learning constitute achievements; in fact, they are among the important achievements of the high school. Therefore we need measures of the effectiveness of pupils in different types of learning, as well as measures of their achievements in school subjects and of their general intelligence. The evaluation of the efficiency of a school is frequently called a survey.

The evaluation of the efficiency of the school is a prerequisite for the most effective kind of supervision of instruction. Before the supervisor can intelligently prescribe changes in the instruction or in the organization of the school, he must know concerning the efficiency of the present procedure. This applies to the entire school system or to any unit of it.

Rating of teachers. The rating of teachers is based, in part, on the achievements of their pupils. However, it is necessary to take into account the quality of the pupil-material with which a teacher is working before we can form any reliable conclusions with reference to a teacher's efficiency as an instructor. A teacher who is working with inferior pupil-material cannot be expected to engender superior achievements. Therefore measures of both general intelli-

gence and achievements of the pupils must be considered in rating teachers. It should be noted that the formal instruction given by a teacher is not the sole source of his contribution to the child's education. The teacher's character, philosophy of life, enthusiasm for his work, etc., have a profound influence upon many pupils. For this reason it is necessary to bear in mind that the rating of teachers should not be based solely upon measures of achievement and general intelligence.

Reports to patrons. It is a custom of our schools to send, at stated intervals, a report of the pupil's progress to his parents or guardian. The entries on these report cards are, for the most part, measures of the pupil's achievements. It is, therefore, obvious that this school activity requires the measurement of the achievements of pupils. These measures of achievement could be more rationally interpreted if they were accompanied by a measure of the pupil's intelligence. There are, however, certain reasons why it would not be appropriate to include this information on report cards.

2. Instructional activities

Diagnosis of pupils with reference to achievements. In considering this activity of the school it will be helpful to recognize two types of instruction. General instruction is based upon general principles. It is applied to all members of a group alike. The same assignment is given to all; all listen to the same explanations. The group is treated as a unit. Pupils are different with respect to their past experiences, and also with respect to their capacities to learn. Hence, some pupils will fail to find in this general instruction the assistance which they need in their learning; others will find this instruction sufficient for their needs.

Supplementary to this general instruction, there is need

for a type of instruction which is based upon the specific instructional needs of individual pupils. This type of instruction may be designated as remedial. An obvious prerequisite is the specific measurement of a pupil's achievements. Such measurement is for the purpose of picking out the pupils who have failed to achieve and, so far as possible, for determining the particular items in which the pupil has failed to achieve. Measurement for this purpose is spoken of as diagnostic.

In order that the measures of achievement may be properly interpreted, it is necessary to have, also, measures of the general intelligence of the pupil. If a pupil has a large capacity to learn, his achievements should be correspondingly high. On the other hand, if he possesses only a limited capacity to learn, we should not expect him to achieve highly. In other words, a pupil's achievements may be expected to be commensurate with his capacity to learn. It is only when a pupil who possesses the necessary capacity has failed to learn that a teacher may expect to find his efforts in the direction of remedial instruction rewarded. If a pupil has achieved up to his capacity to learn, remedial instruction is likely to be fruitless.

Opportunity for diagnosis varies with the grades. The possibility of diagnosing pupils with reference to achievements is not the same in all grades. In the first place, pupils cannot be diagnosed with reference to their achievements until they have had an opportunity to achieve. They must have received some instruction on a topic before diagnosis is possible.¹ In the elementary school the pupils pursue a number of subjects over a period of several years. For example, they study reading in all grades. By repeated drill

¹ This instruction may be incidental. For example, in spelling, a diagnostic test may be given before the pupils have formally studied a given list of words.

they are trained to be skillful in the activity of reading. Much the same situation exists in spelling, handwriting, arithmetic, and, to a less extent, in history and geography. In the fields of these subjects there is an abundant opportunity for diagnosis with respect to achievement. After the pupils have received some instruction, a diagnosis with respect to achievement will be valuable in guiding the teacher in the future instruction.

In the high school, and in some subjects which are taught in the upper grades of the elementary school, the situation is materially different. When a topic has been studied, the pupil does not return to it, except incidentally or in the course of review. There are some topics, such as the operations of algebra, pronunciation of a foreign language, and a few other topics, in which the engendering of skills extends over several months or even a longer period. However, for the most part, the high-school pupil is engaged in the study of topics on which he does not receive continued training. Hence the diagnosis of pupils with respect to their achievements is impossible until instruction on the topics has practically been completed. This condition places very obvious limitations upon the usefulness of measurements of achievement for the purpose of diagnosis in the high school.

Varies also with content of instruction. Another point to be considered is that in the elementary school we have agreed upon certain minimum essentials as educational objectives. Pupils are to be taught to spell the words which they will use in written composition. They are to be taught the operations of arithmetic. They are to be taught to read, and so forth. Although there is not complete agreement with reference to the details of these objectives, it is true that there is a far greater consensus of opinion with reference to objectives in the "tool" subjects of the elementary school

than there is in the case of most subjects taught in the high school. In history, for example, authorities do not agree upon the content of the subject except in the case of a few of the most formal items. In fact, it does not appear to be essential that the content of such a subject as history should be fixed to the extent that the content of handwriting, the operations of arithmetic, or the words of spelling should be. It appears likely that two teachers of European history might be equally efficient in realizing our ultimate educational objectives, but vary widely with reference to the emphasis which they place upon different topics. Indeed, it is conceivable that they might not agree completely with reference to the topics to be studied.

The fact that much of the content of the subjects studied in the seventh and eighth grades and in the high school is not fixed makes it impossible to construct satisfactory instruments for measuring the achievements of pupils. A test which would measure the outcomes recognized by some teachers as minimum essentials would be criticized by other teachers as failing to measure other outcomes which they considered essential. Until there is a greater degree of agreement concerning the minimum essentials of the subjects taught in these grades, it will not be possible to construct standardized objective tests which can be recommended for general use in diagnosing pupils with respect to their achievements. This condition does not remove the need for diagnosis, but it should be made by instruments and methods which are adapted to the instruction which the pupils have received.

Finally, many of the outcomes of instruction in the seventh and eighth grades and in the high school are of such a nature that the problem of measurement is distinctly more difficult than for the more formal outcomes in the elementary school. We have not yet been able to devise

instruments for measuring these outcomes of instruction which yield results that are as satisfactory as those obtained for the more formal subjects. This is particularly true in the case of diagnostic tests.

Thus, it appears that the diagnosis of pupils with reference to achievement in the field of content subjects by means of uniform instruments is limited with respect to opportunity and usefulness. Diagnosis is needed, but until we have more completely agreed upon the specific educational objectives to be attained in content subjects and consequently attempt to produce a uniform product, each teacher must diagnose his pupils, for the most part, by means of methods and instruments which may be adapted to the outcomes he is endeavoring to produce. Furthermore, instruction in these subjects is at present so organized that teachers would be able to make only a very limited use of diagnostic information, even if it were available. Hence, in diagnosing pupils with reference to achievement, the demand for measurement by means of uniform standardized tests is confined largely to the "tool" subjects of the elementary school. At the present time this activity makes little demand for such instruments in the high school.

Diagnosis of pupils with respect to study procedures. The work of the school requires of the pupil a number of different types of study. For example, in handwriting, in the operations of arithmetic, in reading, and in similar subjects, he is asked to acquire skill so that he can do exercises with a high degree of fluency. In other subjects he is asked to solve problems and to acquire abstract and general meanings. In still others the pupil is expected to acquire ideals. The learning activity required of pupils is not the same for the different objectives. Certain study procedures are required for memorizing; others for acquiring ideals. Before pupils can be efficient in their study, they must have ac-

quired a good study technique. They must know how to go about the study of an assignment.

One of the objectives of the school is to train the pupil in methods of learning. In fact, it may be urged that this is the most important objective of the secondary school. The term "preparatory," which for a long time was used to describe the school which the pupil attended before entering college, definitely indicates this function. If a pupil has acquired a good technique of learning, he possesses a very important type of preparation, not only for college, but also for participation in the activities of adult life.

Need for diagnostic study tests. It is obvious that in efficient instruction there is need for the diagnosis of pupils with respect to their effectiveness in the types of learning which the school requires of them. Pupils must be efficient as learners before they can be expected to attain the objectives of the school. The fact that we have just made a beginning in devising instruments for this type of measurement should not be interpreted to mean that the demand for this type of measurement is of minor importance.

In the elementary school the pupil's learning is largely that required for the acquisition of skills. This procedure is relatively simple. It is relatively easy to ascertain by observation whether or not a pupil is following a good procedure. Hence, there is little need for this type of measuring instrument in the elementary school below the seventh grade. In the high school the pupil is required to engage in more complex types of learning. Furthermore, it is not easy to determine by observation whether or not a pupil is engaging in the type of learning which will lead to the solution of a problem or the acquiring of an ideal. There is, therefore, a greater need than in the elementary school for instruments which teachers may use for diagnosing pupils with respect to their study habits.

3. Research activities

Determination of procedure through measurement of outcomes. The scientific attitude applied to education requires that methods and devices of teaching, courses of study, textbooks, and other items of educational procedure be chosen on the basis of their effectiveness in achieving the objectives of education. Therefore, any determination of what procedure is best in education requires the measurement of the outcomes secured by means of the procedures which are being studied. For example, if one wishes to determine the merits of a teaching device, one must apply this device to a group of pupils and measure the outcomes produced by it. Only in this way may we know its worth. Likewise, we can learn the worth of other items of procedure, including types of school organization, only by measuring the achievements of pupils who have been subjected to the procedure.

In any research activity it is necessary to take into account the quality of the pupil-material as well as the character of the achievement secured. Therefore, we have here a demand for both types of measurements. Research is important both in the elementary school and in the high school. The opportunities for research are, however, somewhat greater in the former, because we have more satisfactory measuring instruments for this field.

General summary. The measurement of general intelligence is required in practically all activities in which there is need for any measure of the abilities of pupils. This information is always needed in the interpretation of achievement. It is, likewise, needed for the interpretation of measures of the effectiveness of pupils in the different types of learning. The measurement of general intelligence is needed in all

divisions of the school. It is difficult, if not impossible, to say that the need is greater in certain grades than in others. If the policy of classifying and promoting pupils largely on the basis of their general intelligence is followed, the need for measures of this trait is naturally greater.

The demands of administrative and supervisory activities for the measurement of achievements do not vary with the different grades except as these activities vary in importance. For the reasons which we have pointed out, the usefulness of diagnostic measurement with respect to achievement is limited, for the most part, to the elementary school. On the other hand, the measurement of the effectiveness of pupils in different types of learning has a greater importance in the high school than in the elementary school.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What types of information may we secure about the mental processes of pupils?
2. Distinguish between general intelligence and achievement.
3. Distinguish between general intelligence and specific intelligence.
4. Distinguish between diagnostic measurement and general measurement.
5. What is prognostic measurement?
6. What activities of the school require the measurement of abilities of pupils?
7. What types of measurement are required by each activity?
8. What is the relative importance of measurement in these activities in the different divisions of the school?
9. Give illustrations of the use of measurement in these school activities.
10. Why do we have a much larger number of tests for the elementary-school field than for the high-school field?
11. What are the most important types of tests for use in the elementary school? What are the most important types for use in the high school?
12. Describe remedial instruction. Give illustrations of remedial instruction.
13. What kinds of measurement should be made in a school survey?
14. How are educational measurements related to the work of the supervisor?
15. State the types of measurement most useful in the elementary and in the secondary school fields, and distinguish between them.

SELECTED REFERENCES

BLISS, D. C. "The Application of Standard Measurements to School Administration"; in *Fifteenth Yearbook of the National Society for the Study of Education*, part I, 1918.

BRANSON, ERNEST P. "An Experiment in Arranging High-School Sections on the Basis of General Ability"; in *Journal of Educational Research*, 3: 53-55 (January, 1921).

COURTIS, S. A. "Objective Standards as a Means of Controlling Instruction and Economizing Time"; in *School and Society*, vol. I, pp. 433-36 (March 27, 1915).

COURTIS, S. A. "Supervisory Control by Means of Objective Standards"; in *Proceedings, Second Indiana Conference on Educational Measurements*, 1915, pp. 37-68.

COURTIS, S. A. "Educational Diagnosis"; in *Proceedings, Second Indiana Conference on Educational Measurements*, 1915, pp. 135-71; also in *Educational Administration and Supervision*, vol. I, pp. 89-116 (February, 1915).

COURTIS, S. A. "Courtis Tests in Arithmetic: Value to Superintendents and Teachers"; in *Fifteenth Yearbook of the National Society for the Study of Education*, part I, 1916, pp. 91-106.

DICKSON, VIRGIL E., and NORTON, JOHN A. "The Otis Group Intelligence Scale Applied to the Elementary-School Graduating Classes of Oakland, California"; in *Journal of Educational Research*, 3: 106-15 (February, 1921).

DICKSON, VIRGIL E. "The Use of Group Mental Tests in the Guidance of Eighth-Grade and High-School Pupils"; in *Journal of Educational Research*, 2: 601-10 (October, 1920).

HAGGERTY, M. E. "Some Uses of Educational Measurements"; in *School and Society*, vol. IV, pp. 762-71 (November 18, 1916).

HAGGERTY, M. E. "Measurement and Diagnosis as Aids to Supervision"; in *School and Society*, vol. VI, pp. 271-85 (September 8, 1917).

KIRKPATRICK, E. A. "Measurements, Standards, and Teaching"; in *School and Society*, vol. I, pp. 278-81 (February 20, 1915).

MADSEN, I. N. "Group Intelligence Tests as a Means of Prognosis in High School"; in *Journal of Educational Research*, 3: 43-52 (January, 1921).

MILLER, W. S. "The Administrative Use of Intelligence Tests in the High School"; in *Twenty-first Yearbook of the National Society for the Study of Education*, part II. Bloomington, Illinois: Public School Publishing Company, 1922, pp. 189-222.

PROCTOR, W. M. "The Use of Psychological Tests in the Educational Guidance of High-School Pupils"; in *Journal of Educational Research*, 1: 369-81 (May, 1920).

PROCTOR, W. M. "The Use of Psychological Tests in the Vocational Guidance of High-School Pupils"; in *Journal of Educational Research*, 2: 533-46 (September, 1920).

STARCH, D. "Standard Tests as Aids in the Classification and Promotion of Pupils"; in *Fifteenth Yearbook of the National Society for the Study of Education*, part 1, 1916, pp. 143-48.

STRAYER, G. D. "The Use of Tests and Scales of Measurements in the Administration of Schools"; in *Proceedings of the National Education Association*, 1915, pp. 579-82.

THEISEN, W. W. "The Relative Progress of VII-B Groups Sectioned on the Basis of Ability"; in *Journal of Educational Research*, 5: 295-305 (April, 1922).

TRABUE, M. R. "The Use of Intelligence Tests in Junior High School"; in *Twenty-first Yearbook of the National Society for the Study of Education*, part II. Bloomington, Illinois: Public School Publishing Company, 1922, pp. 169-88.

WHIPPLE, G. M. "The Use of Mental Tests in the School"; in *Fifteenth Yearbook of the National Society for the Study of Education*, part 1, 1916, pp. 149-59.

CHAPTER IV

THE CONSTRUCTION OF EDUCATIONAL TESTS

General structure of educational tests. Practically all educational tests either consist of structural divisions called exercises, or, for the purpose of description, certain divisions are made. We may, therefore, think of the exercise as a structural unit of the test. The different types of exercises and their arrangement produce several types of tests. The test also includes the specifications of the conditions under which the exercises are presented to the pupils.

Types of exercises. Exercises for testing purposes vary in a number of respects. Some of them call for the functioning of the ability being measured under approximately normal conditions; others are highly artificial. Some exercises are easy; other exercises are very difficult. Some require only a few seconds for the pupil to answer; others require several minutes. Some require much writing; others may be answered by making a single mark.

In the operations of arithmetic it is a simple matter to construct an exercise which will call for the functioning of a given ability. Furthermore, the exercise constructed will approximate the normal conditions under which this ability functions. In most cases, if he so desires, the test-maker can easily construct an unlimited number of exercises of this same type which will call for the functioning of the same ability, but which are different in the sense that the answers to the examples will be different. This is the simplest type of exercise construction. It is illustrated by the exercises of such tests as Monroe's Diagnostic Tests in Arithmetic, or the Cleveland Survey Tests in Arithmetic.

Spelling ability normally functions in the spelling of words as they occur in sentences, and when the attention is focused primarily upon formulation of the ideas which the sentences express. Since it is desirable for testing purposes that all pupils spell the same words, it is impossible to construct a set of exercises which will require the spelling of a list of words under completely normal conditions. A spelling test consisting of a list of words to be pronounced separately to the pupils calls for their spelling out of any of context, and the attention of the pupils will be focused primarily upon the spelling. If the test words are embedded in sentences, normal spelling conditions are more nearly, but probably not closely, realized. Hence, in the field of spelling, test exercises must be artificial in the sense that they do not closely approximate the conditions under which spelling ability normally functions.

Devising tests to give an objective record. In some school subjects the normal functioning of the ability does not produce an objective record. In such cases it is necessary to devise an artificial exercise which requires the functioning of the ability to be measured and which will give an objective record. For example, the results of the process of silent reading under normal conditions are not observable. Certain results, such as posture, facial expression, and the functioning of some of the vital organs may be observed in certain cases, but such performances are not satisfactory for testing purposes. In testing silent-reading ability it therefore becomes necessary to devise artificial exercises which require not only silent reading, but also other abilities whose functioning will yield a performance that may be observed and conveniently described. One such type of silent-reading exercise is to have the pupil read a paragraph and then answer a question on it. Another type is to have the pupil reproduce what he has read. In both of these types the pupil is asked

to exercise abilities other than those involved in silent reading.

The examination of available educational tests reveals a large variety of types of exercises. In school subjects where the usual classroom exercises are not suitable for testing purposes, test-makers have exhibited much ingenuity in inventing types of exercises which are suitable for testing purposes and which at the same time satisfy other requirements of test construction. A number of typical exercises from representative silent-reading tests are reproduced below to illustrate what is meant.

MONROE'S STANDARDIZED SILENT READING TESTS REVISED

At every turn the maples burn,
The quail is whistling free.
The partridge whirrs and the frosted burrs
Are dropping for you and me.

What season of the year does the stanza tell about? Draw a line under the one you think.

spring summer autumn winter

In front the purple mountains were rising up, a distant wall. Cool snow gleamed upon the summits. Our horses suffered bitterly for water. Five hours we had ridden through all that arid waste without a pause.

What kind of a country had these people been riding through?
mountainous swampy desert forest valley

THORNDIKE-McCALL READING SCALE

Read this and then write the answers. Read it again if you need to

According to the *Kansas City Star*, the wheat farmers of Kansas are too prosperous to trouble themselves about careful harvesting. They do not cut the fields clean. A gleaner 80 years old,

after the wheat harvest in Pawnee County last year, went over the wheat-fields with a wagon, a rake, a brush, and a shovel and swept up the wheat left on the ground by the threshers. He gathered nine hundred bushels in forty days and sold it at a dollar a bushel.

Do you think most of the wheat which the farmers grew was left on the ground by the threshers?

Why are the fields not cut clean?

.....
Where did the gleaner spend the forty days?

HAGGERTY READING EXAMINATION, SIGMA 3

The champions were therefore prohibited to thrust with the sword, and were confined to striking. A knight, it was announced, might use a mace or battle-axe at pleasure, but the dagger was a prohibited weapon. A knight unhorsed might renew the fight on foot with any other on the opposite side in the same predicament; but mounted horsemen were in that case forbidden to assail him. When any knight could force his antagonist to the extremity of the lists, so as to touch the palisade with his person or arms, such opponent was obliged to yield himself vanquished, and his armor and horse were placed at the disposal of the conqueror. A knight thus overcome was not permitted to take further share in the combat. If any combatant was struck down, and unable to recover his feet, his squire or page might enter the lists and drag his master out of the press; but in that case the knight was adjudged vanquished, and his arms and horse declared forfeited.

1. Underline the word which names the weapon that could not be used:

sword
mace
dagger
battle-axe

2. Check the one of these statements which is false:
 - a. — A knight could fight on foot.
 - b. — One knight could not injure another knight.
 - c. — Mounted horsemen could fight only mounted horsemen.
3. Check the false statements:
 - a. — A knight could be vanquished without being killed.

- b. — A knight's page could fight.
- c. — A vanquished knight retained his horse.
- 4. Check the true statements:
 - a. — Champions were prohibited to use the sword.
 - b. — An unhorsed knight could renew the fight.
 - c. — An opponent was vanquished if his arms touched the palisade.
 - d. — A knight dragged from the lists by his page was beaten.

VAN WAGENEN READING SCALES, GENERAL HISTORY SCALE A

Read the paragraph carefully. Then read the statements below it and put a check mark (V) on the dotted line in front of each statement which contains an idea that is in the paragraph or can be derived from it. The paragraph and the statements may be re-read as often as it is necessary.

The factory system, introduced in America at the beginning of the century, was well developed by the time of Jackson's presidency. The motive power to run the machinery was obtained almost entirely from the swift currents. To accommodate themselves to the new principles of industrial organization, the working classes found it necessary to lay aside the small domestic manufacturing which they had previously carried on in their scattered and isolated homes, and to gather themselves under a common roof, for common effort. A new system of labor was evolved, for the concentration of artisans meant the appearance in the community of a unique social class, possessed of its own special needs, which were different from those of any other class. The factory owners soon realized the exigencies of the new situation, and erected both boarding-houses for the accommodation and protection of the young women in their employ and separate tenements for the married employees and their families, while at the same time they made generous contributions for the support of the schools and churches in the community. The employees were frequently paid, in whole or in part, in "orders" on the company store, where commodities for their use were on sale. A time-table of the Lowell mills for the year 1852 shows that in the month of June of that year the first bell of the factory awakened the operatives at four-thirty in the morning, a second bell came at four-fifty, and the day's work

began by the third bell, early enough to allow of a work period of an hour or so before breakfast.

-1. Previous to the introduction of the factory system manufacturing was done in the homes.
-2. Previous to the introduction of the factory system the American homes were scattered rather than in large groups.
-3. Jackson was President of the United States at the time the factory system was first introduced.
-4. The first kind of power used in American factories was obtained from the rivers.
-5. The factory system was introduced into America about 1700.

Definition of difficulty. Difficulty has been defined as that characteristic of an exercise which, when present in a large degree, causes a large per cent of incorrect responses, and, when present in a small degree, is accompanied by a small per cent of errors. In other words, the degree of difficulty of an exercise is determined by the per cent of correct responses obtained when it is given to a large number of pupils.¹ If certain assumptions are made concerning the distribution of the ability of the pupils to whom an exercise is given and the zero point is located, the degree of difficulty of the exercise can be expressed in terms of a unit of variability of this distribution of ability. This unit is the difference in difficulty between two exercises which are answered correctly by certain per cents of a given group of pupils. The median deviation (*P.E.*) is frequently used as a unit. It is defined as the difference in difficulty between an exercise which is answered correctly by 50 per cent of the pupils, and an exercise which is answered correctly by 25 per cent of the same pupils. The standard deviation (*S.D.* or σ) is also used as a unit. It is the difference in difficulty between an exercise answered cor-

¹ The measure of accuracy in terms of per cent is with reference to standards which may or may not be the same. See page 110.

rectly by 50 per cent of the pupils and an exercise answered correctly by 15.87 per cent of the same pupils. Thus, we describe the difficulty of exercises as being 2.7 *P.E.*, 6.3 *P.E.*, 5.2 σ , etc. The method of determining the difficulty of an exercise will be described on page 95.

Types of tests produced by difficulty of exercises. When the exercises of a test are equal with respect to difficulty, the performance is uniform and the test may be called a *uniform test*. The Courtis Standard Research Tests, Series B, is an illustration of this type of test.

If the exercises of a test vary in difficulty, and are arranged in ascending order of difficulty, we have a *scaled test*. Usually, in a test of this type an effort is made to have a wide range of difficulty represented, and to have the differences between successive exercises equal. Hence, it is necessary to consider the difficulty when selecting exercises for such a test. The Woody Arithmetic Exercises are typical of this kind of measuring instrument. Instead of a single exercise on each level of difficulty, we may have a group of exercises which constitute a uniform test. In fact, this structure is implied when a scaled test is constructed for individual testing. The Thorndike Visual Vocabulary Scales illustrate this type of structure.

If, instead of the arrangement in ascending order of difficulty, the exercises are arranged without regard to difficulty, we have an *irregular test*. In tests of this type, extremely easy and extremely difficult exercises are usually not included, and the exercises of the test are selected on some basis other than difficulty. If the irregularities are relatively small, the test may be treated as being uniform without introducing serious error. This type of test may be illustrated by Charters' Diagnostic Language and Grammar Test, and by Monroe's Standardized Reasoning Tests in Arithmetic.

A *cycle test* is secured when different types of exercises

recur in the test at regular intervals. The exercises of a given type are assumed to be equal in difficulty or approximately so. The Illinois Standardized Algebra Tests represent this principle of test construction.

The name, *spiral test*, has been applied to a modification of the scaled test. A spiral test consists of a number of sub-tests. The exercises within each sub-test are uniform with respect to difficulty or, at most, moderately irregular. The sub-tests differ from each other in difficulty, but they are usually selected on the basis of content rather than difficulty. Generally, the scores of the sub-tests are kept separate. Hence, most of our spiral tests incorporate the characteristics of a scale only in a very crude form. They are essentially nothing more than a battery of uniform tests. The term, *spiral test*, was first applied to the Cleveland Survey Tests. :

Power tests and rate tests. The name, *power test*, has been applied to scaled tests which yield measures of ability in terms of the difficulty of the exercise or group of exercises done with a specified degree of excellence. For this type of test pupils are allowed to continue their work until they are unable to do any more exercises correctly. The Thorndike Visual Vocabulary Scale and the Van Wagenen Reading Scales are examples of power tests. With only a slight modification of meaning, the name, *power test*, could be applied to other types of tests when unlimited time is allowed. The measure of ability would then be in terms of the quality or accuracy of the performance.

A *rate test* is defined as one which yields a measure of the pupil's rate of work. Frequently a measure of the quality of the work is also secured. To measure the rate of work it is necessary either to set a time limit, such that practically no pupil can complete all of the exercises, or to time each pupil separately. In a few rate tests the pupils are allowed to

complete the test after they have indicated the place reached at the end of the time interval allowed. Rate tests are illustrated by the Courtis Standard Research Tests, Series B. A time limit has been applied to all types of tests, but the measure of rate which is obtained has little meaning except for uniform tests, or irregular and cycle tests which approximate uniformity.

Requirements governing the construction of educational tests. It is highly desirable, for practical reasons, that the administration of a test be simple and that it not require a large expenditure of time, but the prime requisite is that it yield scores which are true indices of the ability or group of abilities which it is designed to measure. It has been urged that when this condition is met the details of the structure of the test and the content of the exercises are matters of minor importance. Unfortunately, it is impossible, or at least extremely difficult, to prove conclusively by direct methods that a test does or does not measure that which it is intended to measure. In some cases partial evidence of the validity of a test can conveniently be secured; but for practically all tests there is need for supporting evidence, such as is furnished by an analysis of the abilities measured, the content of the exercises, and the structure of the test.

The production of inferior measuring instruments will be reduced to a minimum by recognizing certain requirements governing the construction of educational tests. Some of these requirements are implied in the assumptions noted in Chapter II. Others refer to the use of the information which the tests yield. In the list of requirements¹ given below there is some overlapping, but it is thought that these state-

¹ These requirements are not stated as criteria for evaluating a test after it has been constructed. They are "guides" to be followed in the construction of educational tests. The evaluation of a test is discussed in Chapter IX.

ments will prove helpful "guides" to one who is attempting to construct an educational test. An elaboration and discussion of these statements is given in the following pages.

1. The performance should maintain a constant functional relation to the ability or group of abilities being measured. This is a comprehensive requirement. Several of the others stated here may be considered subordinate to it. This requirement will be recognized as implied in the first assumption stated in Chapter II.

2. The test should provide adequate opportunity for all pupils to demonstrate their abilities in the field defined by its function.

3. In tests designed to measure school achievement, the performance should be consistent with recognized educational objectives.

4. The performance should be one that is conveniently secured, recorded, and described.

5. The test should provide for controlling the conditions under which the performance is given. This control should be such that the same testing conditions will be approximated by different examiners in different places and at different times.

6. The test should make possible the description of a pupil's performance in terms of the dimensions or characteristics that are significant.

7. In constructing a scaled test, the exercises or groups of exercises should be equally spaced upon the scale of difficulty.

It is not always possible to meet completely these requirements. Frequently, it is necessary to effect a compromise. However, they should be thought of as "guides" in the construction of educational tests. In general, the more completely these requirements are met the better the resulting test will be.

Nature of ability. Although the available analyses of ability are unsatisfactory, our best information indicates that there is not one ability in a subject-matter field, but many abilities which are relatively separate and distinct. At least, they may be so considered for practical purposes. They may involve a general factor, as some authorities contend, but it appears that they include at least some elements which are specific. For example, there may be general factors included in the abilities required for the doing of different types of examples in the field of arithmetic. It, however, appears that the ability to do one type of example is not necessarily accompanied by equivalent ability to do another type of example. Therefore, there are at least some elements of abilities which are not general, but specific. .

The development of ability. Ability develops in two ways. For example, the child learns to read simple material. Practice tends to make him more fluent in this process. He is able to increase his rate of reading and, to some extent, the degree of his comprehension. This is one way in which silent-reading ability develops. The pupil also increases his ability to read by becoming able to read more and more difficult material. Thus, we need to recognize two types of growth, one in the direction of fluency of functioning of certain abilities, and the other in the direction of the development of abilities to do more and more difficult things of the same general type. Educational tests are designed to measure the growth or development of abilities. Some tests have for their purpose the measurement of a pupil's fluency in a given field; others are used to determine the range of a pupil's ability. Power tests belong to the latter class.

Relation of performance to ability. The performance which a pupil gives in response to a test depends upon a number of factors. The abilities which the exercises of the test call for may be limited to the ones which we are attempting

to measure. In this case we have direct measurement. Most educational measurements are, however, semi-direct or indirect, and the abilities being measured are combined with others required by the test. For example, in the measurement of silent-reading ability, it is necessary to require the pupil to reproduce what he has read, answer questions, or give some other performance that can be observed. The functioning of abilities other than those involved in the process of silent reading is required to produce such performances. It has recently been shown that a performance consisting of the writing of the sums of the fundamental addition combinations depends upon the ability to write figures, as well as upon a knowledge of the number combinations.¹ It was found that some pupils were limited in the scores which they were able to make by their slow rate of writing the answers to the number combinations. Had they been able to write more rapidly their scores would have been materially increased. On the other hand, there were pupils who were able to write more rapidly than they could give the number combinations. Since the rate of writing figures is variable from pupil to pupil, the performance is a function of the rate of writing figures as well as of the ability to recall the fundamental addition combination.²

The performance is also affected by a number of other factors, such as the effort which a pupil makes, his physical condition, his emotional status, the time of day, the form in which the test is presented, the manner of the examiner, the explanations and directions given to the pupil, acquaintance

¹ Courtis, S. A., and Thorndike, E. L. "Correction Formulæ for Addition Tests"; in *Teachers College Record*, vol. XXI, pp. 1-24 (January, 1920).

² The investigators propose a correction formula by means of which the performance which a pupil would make if he were not handicapped by the writing of the answers can be calculated. Such a correction formula is useful for experimental purposes, but it introduces a procedure which is too complex for general use.

with testing procedure and the particular type of exercises used, specific coaching for the test, the time allowed, and other more subtle factors. The performance is, therefore, a function of a number of variables. The functional dependence of the performance can be represented in mathematical terms by the following equation:

$$P = f(a_1, a_2, a_3, \dots a_n, x_1, x_2, x_3, \dots x_n).$$

In this equation, the a 's represent the abilities to be measured and the x 's all of the other factors which affect the performance.

1. Requirements for constant functional relationship

How to secure this. A constant functional relationship between the performance and the abilities being measured means that the same relation exists for all pupils, even when tested by different examiners and at different times and at different places. The necessity for this constancy of relationship is due to the use of the same set of norms for interpreting the scores of all pupils. The first step in securing a constancy of functional relation is to specify the abilities to be measured. Our purpose may be to secure a general or average measure of the abilities within the entire field of a school subject, or even within a large section of this field. On the other hand, we may desire a measure of the abilities within a very restricted field, such as single-column addition of integers, seven figures to the column. Whatever field of abilities is chosen, the exercises should be limited to that field and should be representative of it. In the case of general measurement, this requires that they form a random sample of it. The use of a variety of types of exercises within a single test will result in the score being an unanalyzed total, unless the test is sufficiently long and the types are distributed in a random way. In such a case, the score

may be looked upon as an average or general measure of the abilities in the field represented by the exercises of the test.

The abilities introduced for testing purposes, of course, are variable from pupil to pupil. The other *x*-factors also exhibit variable tendencies unless proper precautions are taken. In order that the performance shall approximate a constant functional relation to the abilities measured, it is necessary that the effect of other abilities be eliminated or reduced to a minimum, and that the other *x*-factors be controlled so that there will be a minimum of variation from pupil to pupil. This ideal can only be approximated, but the test-maker should recognize the importance of the control or elimination of all variable factors except the abilities being measured.

Excluding other abilities than those it is desired to test. The exercises should be constructed so that abilities other than those being measured will be involved to the smallest extent consistent with other requirements of test construction. For example, exercises should be constructed so that a minimum of writing is required. When possible, it should be reduced to checking or marking certain words or statements. The Burgess Picture Supplement Scale for Measuring Ability in Silent Reading illustrates the introduction of supplementary abilities that tend to destroy the constant functional relationship between the abilities measured and the performance. The pupil is asked to draw pictures. Although only very simple drawings are called for, all the pupils do not interpret this request in the same way. Some try to produce a performance which possesses merit as a drawing. Others make only very crude pictures. As a result, a variable factor is introduced. General intelligence tests which require reading introduce a variable ability in the primary grades, since some of the pupils have not learned to read.

Direct measurement is to be preferred whenever it is feasible. In this case, the exercises call only for the functioning of the abilities to be measured. However, it is seldom possible to realize direct measurement. Most educational measurements are semi-direct. A semi-direct performance may bear a constant relation to the ability being measured, but such a relation cannot be assumed. Its existence must be demonstrated experimentally. In a few tests, the abilities which function in the production of the performance do not include those which we desire to measure, and the measurement is called indirect. It is, however, possible that, even in the case of such indirect measurement, a constant functional relation may exist between the performance and the abilities which it is desired to measure. If it exists, it is a coincidence and must be demonstrated experimentally.

Controlling other x -factors. The other x -factors are controlled by the general structure of the exercises and test, and by the directions which are devised for its administration. Since tests are designed to be used by different examiners at different times, it is, of course, impossible to devise a measuring instrument which will be absolutely "fool proof." Examiners who are not in sympathy with the use of educational tests or who are inclined to be careless will, as a result, introduce variable factors. But it is possible to formulate directions to examiners sufficiently complete and sufficiently clear so that examiners who wish to control such variable factors as explanations, timing, and manner of presenting the test, will be successful to a large extent. Control of these factors will, in turn, tend to control the effort the pupil makes and his emotional status. When pupils are not accustomed to educational tests, or the exercises are unusual in structure, second-trial scores will be materially larger than first-trial scores, especially in the case of rate tests. Third-trial scores will be slightly in excess of second-

trial scores, but the frequent use of educational tests tends to reduce the differences between the scores yielded by successive trials. The effect of acquaintance with the type of exercises can be partially eliminated by the use of a fore-exercise, or preliminary test. The use of a printed copy of the test for each pupil makes the form of presentation more uniform for all pupils. It is not in the power of the test-maker to eliminate deliberate coaching.

2. Requirement for unrestricted functioning

The type of ability to be measured. If the performance is to be a valid index of the ability of pupils it is obvious that each pupil to whom the test is administered must be given equal opportunity to give expression to his abilities within the field of the test. A test-maker needs to be acquainted with the nature of the abilities he is attempting to measure. If he has to deal with a general or average ability he should proceed differently from what he would do if he were devising an instrument to measure separately a number of specific abilities. He must, also, be guided by the type of mental growth he desires to measure. It is also necessary that the pupil be afforded an opportunity to produce a performance which will reflect the significant characteristics of the ability being measured. For example, if the rate of functioning is a significant characteristic of this ability, the pupil must be given an opportunity to demonstrate his rate of work.

In certain types of silent reading the rate of reading is a significant characteristic of the pupil's ability. Therefore, tests designed to measure such types of silent-reading ability must provide an opportunity for the pupil to demonstrate his rate of reading. For certain abilities the quality or accuracy of the performance is the significant characteristic. In other cases the level of difficulty on which the pupil is just barely able to work with a specified quality of performance is

a significant characteristic.¹ Therefore, the type of test to be constructed depends upon the nature of the abilities to be measured and the purpose for which the test is used. A performance must be large enough to represent adequately a pupil's ability. A single exercise is not sufficient. When a measure of the ability of individual pupils is desired, each pupil must be given the opportunity to do several exercises which call for the same ability. Except in extreme cases, the longer the performance the more truthfully it will represent a pupil's ability.

Limitations of functioning due to content. The content of the test places certain restrictions on the opportunity which the pupil has to demonstrate his abilities. He has no opportunity to demonstrate any except those called for by the test. If it is not possible to include exercises which call for all of the abilities we desire to measure, the test should be representative of these abilities. For example, if the purpose of a spelling test is to obtain a measure of the pupil's ability to spell the most frequently used words of the English language, the test words should be representative of the most frequently used words of the English language. If the function of an arithmetic test is to yield measures of a pupil's ability to do addition examples, the test should consist only of addition examples, and these should be representative of the various types of addition examples which exist.

Limitations of functioning due to the structure of a test. Certain limitations are imposed in the construction of a test by the type of structure of the test itself. These may be enumerated under the following headings:

1. *Uniform performances.* Most uniform tests consist of exercises similar in structure and content. This is the case

¹ Difficulty is a characteristic of the thresholds of ability. It tells just how far the pupil is able to go in the doing of more and more difficult tasks of a given kind with a specific standard of accuracy.

in the Courtis Standard Research Tests, Series B. Each test of this series is restricted to a single type of example in one of the four fundamental operations. The pupil is given no opportunity to demonstrate his ability to do other types of examples. He is able to demonstrate his ability to do only the one type of example. This constitutes a limitation of the uniform performance. On the other hand, most tests of this type are timed so that a measure of the pupil's rate of work is secured. Thus, the pupil is restricted by the scope of the test, but is given an opportunity to demonstrate his fluency within a limited field.

2. *Scaled performances.* A scaled performance test presents a variety of exercises to the pupil. Those on different levels of difficulty generally differ in structure and content as well as in difficulty. Thus, the pupil has the opportunity to demonstrate the range of abilities which he possesses within the field of the test. In this respect, a scaled performance test is superior to a uniform test. On the other hand, a scaled performance test is not usually timed. When there is a time limit it is generally set so that practically all pupils have ample time to do all of the exercises which they are able to do correctly. This is equivalent to no time limit at all. In a few cases the time limit is set so that practically no pupils can complete all of the exercises; but this practice is confined for the most part to tests of general intelligence. In Gray's Oral-Reading Test the testing is individual, and the pupil is timed on each level of difficulty. However, generally, the pupil has no opportunity to demonstrate his rate of work on a scaled performance test. This is a serious limitation of this type of test, when the rate of work is an important characteristic of a pupil's ability.

The exercises for a scaled test are selected because they are found to possess a certain degree of difficulty, rather than because they represent certain content. It may, therefore,

happen that certain types of exercises are not found in a scaled test because they were not found to possess the exact degree of difficulty desired. It appears that this happened in the construction of Woody's Arithmetic Exercises, and certain important types of examples are not found in these tests. Thus, the pupil is deprived of the opportunity to demonstrate his ability to do these examples.

In a number of scaled performance tests the pupil is given only one exercise on each level of difficulty. One trial does not afford the pupil an adequate opportunity to demonstrate his ability. He either does the one exercise correctly or fails completely.¹ In such a case no reliable index is obtained of the pupil's ability to do exercises on a given level. It may be noted that this restriction does not apply if the test is used to secure an average measure of the ability of a class or larger group. If each of thirty pupils does an exercise we have the same amount of data as we have in the measurement of the ability of a single pupil when he does thirty exercises. In a few scaled performance tests the pupil is given more than one exercise on each level of difficulty. For example, the Thorndike Visual Vocabulary Scale presents the pupil with ten words on each level of difficulty. In other cases, the test-makers have provided for computing a pupil's score from his performance on several levels of difficulty. If it is assumed that the ability measured is general, or that only an average measure is desired, this procedure tends to compensate for not giving the pupil an adequate opportunity for demonstrating his ability for doing exercises on each level of difficulty.

3. *Spiral performances.* A spiral performance possesses the virtues of both the uniform and the scaled performance. The sub-tests may be, and usually are, timed. Hence, the

¹ A few scaled performance tests provide for partial credit for exercises partly correct.

pupil is given the opportunity to demonstrate his fluency on each level of difficulty, and the sequence of levels provides the opportunity for demonstrating the range of his abilities. If the levels of difficulty are chosen because they call for abilities that are educationally worth while, and not merely for statistical reasons, the exercises may be made representative of the field. The number of levels of difficulty that are necessary to represent satisfactorily the abilities in a field depends upon the number and importance of the specific abilities in the field. However, time and expense, and the difficulties of describing the performances, limit both the number of levels and the number of exercises on each level.

4. *Irregular performances.* The irregular performance test is used because satisfactory uniform exercises are not available, or because it is deemed advisable to have certain content represented. This type of test may be constructed so that the pupil is given the opportunity to demonstrate his ability over a considerable range, but, since we do not have any satisfactory plan for describing irregular performances, this opportunity is limited in significance. Irregular tests may be timed; but, since the exercises are not uniform, and generally not confined to one topic, the pupil is handicapped in demonstrating his fluency. When the irregularities are small, the test may be treated as uniform without introducing a serious error. Monroe's Standardized Silent Reading Test, Revised, is an illustration of a slightly irregular test which is timed.

5. *Cycle performances.* The cycle performance is a device to approximate certain of the good features of a uniform performance, and at the same time allow for the expression of a group of abilities rather than of a single ability. This device has its greatest use in dealing with a large group of important abilities when it is necessary to economize time. Of course, the resulting description will permit only general

interpretation. Cycle tests are usually timed. Thus the pupil has the opportunity to demonstrate his fluency.

Exercises must not be ambiguous and must be explicit in their specifications. If an exercise is not interpreted alike by all pupils they will not have equal opportunities to demonstrate their ability. If the specifications are indefinite, different pupils will attempt to give different kinds of performances. One limitation of questions in which pupils are asked to "discuss" or "compare" is the indefiniteness of the demands of the questions. In order to be satisfactory for testing purposes an exercise should make clear to the pupil exactly what he is to do.

3. Agreement with educational objectives

Should measure the effectiveness of the instruction. Most achievement tests have for their purpose the measurement of abilities which have been engendered in the pupils by the instruction of the school. This means that the abilities which such a test measures should be included in the objectives of instruction. Otherwise, the measurements secured by using it will not be a true index of the effectiveness of the instruction. Furthermore, the scores cannot be used as a basis for planning remedial instruction. For example, if a teacher has organized her instruction in arithmetic primarily for the purpose of attaining certain objectives in the solving of problems, the measures resulting from the use of a series of tests upon the operations of arithmetic will be of little value as an index of the effectiveness of this instruction.

The use of an achievement test, particularly for the purpose of diagnosis, suggests to both the teacher and the pupils that the norms for the test are worthy educational objectives. In fact, the use of test norms as educational objectives has been generally urged. When the norms are used in this

way, it is imperative that the content of the exercises and the structure of the test should be in agreement with our educational objectives. This statement applies only to achievement tests. In the case of instruments for measuring general intelligence, the restriction does not apply, because the norms for these tests are not intended to be considered as educational objectives.

Exception in the case of a school survey. For certain purposes, such as a school survey, it may be appropriate to use tests whose exercises and structure are not entirely consistent with recognized educational objectives. In fact, the position may be taken that the most desirable test in such a case would be one which discovered what the pupils had learned. For survey purposes, it may be important to know whether the pupils have acquired abilities which are not in agreement with accepted educational objectives. When such a condition is found to exist, it indicates an improper distribution of teaching time or emphasis. For example, if a group of pupils were found to have been trained in doing very difficult and intricate exercises in arithmetic or algebra, one would be justified in criticizing the school unless it could be demonstrated that the achievements were socially worthwhile. This, however, is a special use of educational tests, and for this reason the position is taken that in general the content of the exercises and the structure of educational tests should be consistent with recognized educational objectives.

Prior determination of objectives. At the present time our educational objectives have been adequately determined for only a few of our school subjects, such as spelling, handwriting, and arithmetic (in part). In many subjects only general statements of our objectives are available. Hence, it has frequently been necessary for the maker of a test to conduct first a rather elaborate investigation to obtain in-

formation with reference to these matters as a preliminary step in the construction of his test.¹ This preliminary step does not always appear in explicit form in the accounts of test construction. This is because it is assumed that the abilities which the test is designed to measure are included in our objectives. For example, our present practice clearly indicates that the abilities required in performing the operations in arithmetic with integers are included in our objectives in arithmetic. No extended investigation is required to prove this. Therefore, the test-maker in this field may proceed at once to construct tests to measure these abilities.²

In other fields, where our objectives are not so obvious, it has been necessary to determine what they are before a satisfactory test can be constructed. For example, in the field of spelling, a test which would measure the extent to which pupils were attaining our objectives was not possible until we had defined our objectives. When our objectives in the field of spelling were defined as teaching pupils to spell the most commonly used words of the English language, and a list of these words had been obtained, as in the Ayres Spelling Scale, test-makers were then in a position to construct spelling tests. In the less formal school subjects, such as geography, history, physics, and literature, our objectives have not as yet been defined in the way required for test construction. Until this is done the activities of test-makers in these fields will be decidedly limited.

Relation of a scaled performance test to educational objectives. The idea underlying a scaled performance test is

¹ Ayres, L. P. *A Measuring Scale of Ability in Spelling*. Bulletin of Education, 139. Russell Sage Foundation, New York City.

Monroe, Walter S. *Report of the Division of Educational Tests for 1919-20*. Bulletin No. 5, Bureau of Educational Research, University of Illinois. The derivation of Monroe's Standardized Reasoning Tests is given.

² Certain limitations must be observed. One must not assume that extreme types of examples are included in our objectives.

that the pupil will advance along the scale of difficulty until he reaches a step which he can just do successfully according to the standards of success that have been established for the particular scale under consideration. These standards differ for different tests. When expressed in terms of a per cent of perfect accuracy, 50 per cent, 75 per cent, and 80 per cent have been used. From a statistical point of view 50 per cent is to be preferred. When a pupil's performance is described in terms of the highest level of difficulty reached, his score includes no mention of his performance on other levels of difficulty.

In connection with the use of the norms for scaled performance tests as educational objectives, two criticisms may be made. First, a standard of accuracy which is preferable for statistical reasons may not meet social approval. In arithmetic or spelling, an objective of less than 100 per cent accuracy is open to objection. To hold before the pupil an objective which requires only 50 per cent accuracy, or even 75 per cent accuracy, is contrary to our educational theory and practice. Second, the norms of a scaled performance test suggest that the teacher's objective should be to train the pupils to do more and more difficult exercises because of their degree of difficulty.¹

Such an objective is false for two reasons. First, pupils are not to be trained to do things just because they are difficult, but because they are educationally worth while. A very difficult exercise may have a high educational value, but, in many cases, an exercise is difficult just because it calls for abilities which are not considered an important edu-

¹ This objection to scaled performance tests could be overcome in part by adopting a method of scoring by which the credit given for doing an exercise correctly would be proportional to its educational importance. This would mean that little credit would be given for doing a very difficult exercise correctly if it was unimportant educationally. However, this feature has not yet been incorporated in any of our scaled performance tests.

cational objective, and hence have not received emphasis in the school. This would be true of many questions in history and geography which call for facts of minor importance, and also of complicated exercises in arithmetic and algebra. In the second place, such an objective does not include the idea of training pupils to do things well and rapidly. Training for fluency is not equally important in all fields. For example, it is less important in problem solving in arithmetic than in the operations. Even in the case of silent reading, we may set up objectives which do not include fluency. However, when fluency is included in our objectives, the norms for a scaled performance test will be inconsistent with them. Thus, we may say that, in general, scaled performance tests as they are usually scored are inconsistent with our educational objectives in many fields.

Incidentally, it may be noted that this type of test implies that ability is general rather than specific. It is assumed that the same abilities function in the difficult exercises as in the easy ones.

4. Requirements for the administration of a test, and the description of the performance

Elements in a complete performance. From a practical standpoint it is necessary that the test be one which can conveniently be given. For anything like general use, this means that the test must be such that it can be administered to groups of pupils, rather than only to individual pupils. The time required should not be excessive, and the test itself should be inexpensive. No elaborate accessories should be required for its administration.

The complete description of a performance involves giving the magnitude of its three dimensions:

1. The amount, or the rate at which the ability functioned in producing it.

2. Its quality or accuracy.

3. The type of exercise in response to which it was given.

Under certain conditions, this third dimension may be given in terms of the level of difficulty reached. The description of a performance with reference to its amount has no meaning unless the test is timed. Hence, if it is desirable to have a measure of a pupil's rate of work it is necessary that such a time limit be fixed that no pupil can finish or that the time used by each pupil be ascertained. It is highly desirable that the performance be such that its quality or accuracy can be objectively described. When a scorer is required to exercise judgment, the measures tend to become less accurate. For this reason it is frequently necessary to reject exercises which are otherwise satisfactory because the resulting performances cannot be objectively described. The specific requirements from the standpoint of the description of the performance depend upon the kind of information the test is designed to yield. For a rate test there are certain requirements. For a power test they are different.

The interpretation and use of the scores will be made easier if they are expressed from an absolute zero point and in terms of a constant unit. We shall then be able to say that a score of 24 means twice the ability indicated by a score of 12, and that a difference of 4 obtained from 9 to 5 is equal to the difference of 4 obtained from 31 to 27. These are not necessary requirements, but should be classed as desirable.

5. Control of testing conditions

Testing conditions are controlled when they are made the same for all pupils, even though tested at different times and by different examiners. This has been mentioned briefly in connection with the requirement of a constant functional relationship between the performance and ability. It will,

however, be helpful to emphasize the manner in which this control is to be secured. The test-maker cannot completely control testing conditions, but much can be accomplished by appropriate precautions. Attention is called to the following items:

Attitude of the examiner. The examiner should maintain the attitude that the purpose of giving the test is to secure truthful information concerning the abilities of the pupils. The performance desired is one that is representative of their abilities. If the pupils are excited they are not likely to give a performance that is a truthful indication of their ability. If the examiner approaches them with an attitude of indifference the pupils are likely to become indifferent. If the pupils have received coaching on the particular exercises, their performance will not be a truthful indication of their abilities. By an appropriate explanation of the test and by suggestions the test-maker can do much to engender a proper attitude in the examiner.

Preparations by pupils. The examiner should be directed to have the pupils clear off their desks and provide themselves with well-sharpened pencils, unless they are accustomed to using pen and ink. In the lower grades pencils are preferable. In the upper grades the use of pencils is optional. It is well for each pupil to be supplied with an extra pencil, in case he breaks the point of the one in use. The pupils should have no material on their desks except that which they are to use in the test. If they are to make use of any paper, they should be provided with one or two sheets of plain paper.

Distribution of test papers. Such a simple thing as the distribution of the test papers should not be left to the judgment of the examiner. The pupils should be cautioned not to examine or open the test folders until directed to do so. It is generally necessary to give special emphasis to this di-

rection. The examiner should not distribute the test papers himself. This should be done by two or three members of the class. The distribution will be facilitated if the examiner counts out the number of test papers for each row. It is well to have the tests placed on the desk with the back up. This will lessen the possibility of some pupils beginning work before the signal to begin is given.

General data. Blanks calling for the age, name, and grade of the pupil, and other general data, should be printed in an appropriate place. In case the test is one in which the pupil is giving his performance on plain paper, he should be asked to record this information. The examiner should be cautioned to examine the entries of the pupils, particularly in the case of the younger pupils. This may be done after the test is completed and corrections may be made from the teacher's records. It is very important that the grade and age of the pupil be given correctly. Errors in these items do not affect the accuracy of the scores, but they may lead to erroneous interpretations of the scores.

Explanation of tests to pupils. The nature of the exercises which the pupils are to do must be explained to them. Sometimes this is by means of a simple description. Sometimes it is accomplished by giving them a preliminary exercise of the same general kind. In any case the purpose is to have each pupil understand what is expected of him in the test. The explanation of the nature of the test should include a statement of how the pupil is to do the exercises. For example, in the case of arithmetic, a pupil should be informed whether he is to check each example before proceeding to the next, or whether he is to try each example only once. It is imperative that the same explanation be given by all examiners. Hence it is necessary that the test-maker provide the exact explanation which is to be given, and that each examiner follow closely these instructions.

Timing the performance. In case it is desired to secure a measure of the rate at which a pupil works it is necessary to have a time limit such that no pupil will complete the test. It is also necessary that the time limit specified be rigidly observed in giving the test. All pupils must begin at a given signal, and stop promptly at another signal. It is well to have the pupils assume an attitude of attention, such as raising the right hand and looking at the examiner, preliminary to giving the signal to begin. Requiring that pupils again assume this attitude of attention when the signal to stop work is given will ensure that all pupils stop at the same time.

Illustrative test directions. The following directions for Monroe's Silent-Reading Tests, Revised, are given as typical of the directions for many tests:

GENERAL INSTRUCTIONS

Tell the pupils not to open the test folder until directed to do so. Then have two or three pupils who occupy front seats distribute the folders with the first page up, placing a copy on the desk of each pupil.

Be certain that each pupil is supplied with a well-sharpened pencil. It is better to use pencils than pens, although, if the pupils are more accustomed to the pen, it may be used. Have the blanks at the top of the first page filled in, assisting pupils if necessary. Be certain that none of the items are omitted. *Be careful that the pupil's age on his last birthday and the date of his next birthday are given correctly.*

It is very important that exactly the time specified for each test be allowed for it. An allowance of two minutes means exactly two minutes, not two minutes and five seconds or ten seconds. It may be contended that, if exactly two minutes are allowed, the pupils will not be working two minutes because it will take them a little time to get started. This may be true, but the important thing is that pupils in different schools shall be given the same time allowance. The only way in which this can be accomplished is for all teachers to follow directions exactly.

In order to allow exactly the number of minutes specified for the different tests, it will be necessary for the person administering the tests to have a watch with a second hand. A stop watch is even better, although an ordinary watch that has a second hand will be satisfactory if a little care is exercised in using it. The best plan is for the teacher to notice the position of the second hand when the signal to begin work is given and write down this position. When the second hand reaches this point again, a mark should be made. Do this for each time the second hand reaches this point.

The following instructions should be read to the pupils. In order that they may be read effectively, the teacher should become thoroughly familiar with them before giving the test to the pupils. The portions of the following directions enclosed in parentheses and printed in italics are not to be read to the pupils. They are "stage directions" to the teacher. The directions printed in the test folder are reproduced. Hence, it will be unnecessary for the teacher to refer to it in reading the directions.

DIRECTIONS TO BE READ TO THE PUPILS

Below there are three exercises. Under each exercise there is a row of words printed in bold-faced type. Each exercise asks a question. You are to read each exercise and then answer the question by drawing a line under the right word printed in the black type.

Read the following exercises:

- (a) I am a little dark-skinned girl. I wear a slip of brown buckskin and a pair of soft moccasins. I live in a wigwam. What kind of a girl do you think I am?

Chinese French Indian African Eskimo

The answer to this exercise is "Indian," so draw a line under Indian. (*See that the pupils draw a line under the word Indian in this exercise.*)

- (b) Spring is the time for planting seeds. They grow fastest in summer. Autumn is the harvest time. When are seeds put into the ground?

Spring Summer Autumn Winter

The answer to this exercise is "Spring." Draw a line under Spring.

- (c) In the sunny land of France there lived a sweet, little maid named Piccola. Piccola's father was dead, and her mother was very poor. Draw a line under the word below that tells in what country Piccola lived.

Germany Russia France England

(The correct answer to this exercise is France. Do not tell the pupils what the answer to this exercise is until they have had an opportunity to study it. Before going on, ask the pupils if they understand what they are to do.)

On the three following sheets there are a number of exercises like these to be read and answered. When the signal is given, turn over this page and begin. Work rapidly, but remember that your answers must be right in order to count. Remember that you are to draw a line under only one word in each exercise. Also remember that this test is on three pages. When you finish one page turn to the next.

Turn to the next sheet, but do not begin work. Attention! Pencils up. *(Look at your watch and note the position of the second hand.)* Ready, Go! *(Write down the position of the second hand. Allow exactly four minutes.)* Stop! Attention! Draw a line through the number at the left of the line which you were reading when the signal to stop was given.

6. Description in terms of significant dimensions

This requirement has been discussed incidentally in connection with the agreement of the norms with educational objectives and the description of the performance. There is little to add here. The dimensions that are significant in a particular case depend upon the use that is made of the test. If the test is to be used to secure a diagnosis of pupils in a field where the educational objectives are defined in terms of fluency, the significant dimensions are rate and accuracy, or some combination of them. If, on the other hand, the educational objectives are defined in terms of power or "range of acquaintance," the test should provide for the description of the performance in terms of the level of difficulty reached.

Thus, the test-maker must first define the function of his proposed test. Users of tests should exercise care in selecting tests that are appropriate for their purposes.

The law of the single variable. The author ¹ of a recent test has formulated a principle with reference to the description of performances which she has called *the law of the single variable*. Briefly stated, this law means that the three variables or dimensions must be recognized in describing pupils' performances. If one of these dimensions is not constant for a group of pupils its variation must be recorded for each pupil in the process of testing, and the scores must be interpreted on the basis of the single variable. For example, if the pupils are to be compared with respect to their rate of work it must be shown that both the quality and the difficulty of the work done were the same for all pupils. "All the law of the single variable does not permit is the attempt to compare combinations of the three variables in unknown or varying amounts."

There is some implication in the account of the law of the single variable that it means that all pupils should be forced to give performances which are constant with respect to two of the three variables. It is obvious that this interpretation of the law cannot be justified. The characteristics of the abilities which pupils acquire are not restricted to a single variable. When a group of pupils are performing in a given subject-matter field in the way which is most natural for each, large individual differences are exhibited with respect to these three variables, particularly with respect to rate and quality of work. This is true even in the case of pupils who have received the same instruction. Some will work with emphasis upon rate; others will emphasize accuracy and work more slowly. If a uniform test is used, the

¹ Burgess, May Ayres. *Measurement of Silent-Reading Ability*. Division of Education, Russell Sage Foundation, 1921.

difficulty is constant for all pupils. If a scaled performance test is used, different pupils will be working upon different levels of difficulty after the first few seconds of the time allowed for the test. Since some will advance much farther along the scale than others, the average difficulty of the exercises done will vary from pupil to pupil. In certain subject-matter fields the rate of work is relatively unimportant, and even though all pupils do not work at the same rate, variations in the rate of work may be disregarded. For example, in spelling the rate of work is not considered important. The rate also appears to be unimportant in the case of painting or drawing when the products are real works of art.

If a variable is controlled by an arbitrary procedure which produces unnatural conditions, it appears likely that the ability which functions is modified, or factors are introduced into the testing procedure which produce the same result upon the performance. For example, if in the case of handwriting all the pupils of the group were forced to write at a fixed rate, those who were accustomed to write more rapidly or more slowly than this rate would exhibit abilities different from those which normally function in their handwriting. Therefore, any attempt to control by arbitrary conditions one or more of the dimensions of a pupil's performance is likely to introduce serious errors in the resulting measures.

The real point of this law, as indicated above, is that measuring instruments must explicitly recognize the possible existence of the three variables or dimensions of a pupil's performance. These must be described separately if accurate interpretation of the scores is to be possible. If a variable is omitted in a pupil's score, it must be shown to have been constant in its effect upon the performances of the different pupils or to be socially unimportant. When two

or more of the variables are combined in a single score, this combination is likely to consist of "unknown and varying amounts" of the different variables. For example, the number of exercises done correctly is frequently used as a pupil's score. This is a combination of the rate and the quality of the performance. The combination is in the form of a product, but in no case does the score show the relative magnitude of the two factors. Hence, in interpreting scores which are the number of exercises done correctly, it is necessary to recognize the limitations of an unanalyzed total. It is, however, possible that the practical convenience of having the performance described in terms of a single score largely compensates for the loss in precision of the interpretation.

7. Equal spacing of exercises on the scale of difficulty

If the exercises of a scaled performance test are equally spaced on the scale of difficulty, certain simplified methods of describing the performance may be used. (See page 123.) The equal spacing also gives a more random selection of exercises on the basis of difficulty.

Selection of exercises for a test. The requirements discussed in the preceding pages should guide one in deciding upon the type of test to be constructed, and indirectly in the construction of the exercises. It is now necessary to consider the procedures to be followed in the selection of the exercises required for a particular type of test. This will be discussed under three heads.

1. Content. If a general test is to be constructed, each exercise should call for all of the abilities within this field, or the group of exercises chosen should call for a random sampling of them. For example, a spelling test which is designed to yield a general or average measure of the pupil's spelling ability cannot include all words which the pupil should be

able to spell. Therefore, it should consist of words which have been selected so that they constitute a random sample of the words included in our objective in spelling. The same situation should exist with reference to general tests in other fields, such as geography, history, silent reading, and so forth. One procedure that has been followed in order to secure exercises that constitute a random sampling is to take the subject-matter as presented in a textbook and select items appearing at regular intervals. For example, in the case of spelling, we have had spelling tests constructed by taking the first defined word appearing on the even-numbered pages of a dictionary. The same method has been used for selecting the words for a vocabulary test. Another method which has been followed has been to compare a number of textbooks and to use the items which were common to the several texts. Sometimes this procedure has been supplemented by securing the opinions of competent judges with reference to the importance of the exercises.

One author has selected exercises for a language test on the basis of pupil needs.¹ Performances of pupils given under normal conditions were examined to ascertain the grammatical errors which they made, and exercises were then constructed which called for the correction of the errors occurring most frequently. A somewhat similar method was followed by Ayres in constructing his spelling scale. He selected words occurring most frequently in the correspondence and other writings of adults. They are the words of the English language most frequently used in writing.

2. *Suitableness for testing purposes.* It frequently happens that an exercise will prove on trial to be unsuitable for testing purposes. Except in the construction of a scaled test, very easy exercises and very difficult exercises are to be

¹ Charters, W. W. "The Construction of a Language and Grammar Scale"; in *Journal of Educational Research*, vol. 1, p. 249 (April, 1920).

avoided. All exercises which are confusing or ambiguous should likewise be avoided. Occasionally exercises must be rejected because the performances secured by them cannot be conveniently described. An experienced test-maker will be able to avoid including a large number of unsatisfactory exercises in the preliminary list; but the only certain procedure is to have the exercises given to several hundred pupils and to analyze the responses that are obtained.

3. *Difficulty.* We are concerned with the difficulty of exercises in both uniform and scaled tests. In the former, if the exercises have not been constructed so that equal difficulty may be assumed, it is necessary to have the exercises given to a large number of pupils so that those which are done correctly by the same per cent of the pupils may be selected. In the construction of a scaled performance test it is necessary to select from a preliminary list those exercises which will form a difficulty scale. This requires that the difficulty of each exercise in the preliminary test be determined and expressed in terms of a convenient unit. In doing this, the exercises are given to a large number of pupils, and the per cent of correct responses for each exercise is calculated and translated into the corresponding difficulty value.

A number of test-makers have followed a very elaborate procedure in constructing such a scale, but some recent studies of the scoring of scaled performance tests indicate that serious errors would not be introduced if the procedure were greatly simplified. For example, the difficulty of each exercise might be computed from all grades combined instead of for each grade separately. Another possible procedure is to base the difficulty values upon the results secured from only one grade. Either method will eliminate the calculation of the inter-grade interval, which will greatly shorten the process for tests designed to be used in a sequence of

several grades. We shall describe the complete process, but any one contemplating the construction of a scaled performance test may well consider the advisability of abbreviating the procedure.

Assumptions underlying the construction of scaled tests. The makers of scaled tests have made two assumptions. First, it is assumed that when an unselected group of pupils, such as those belonging to a given school grade, is distributed according to a given ability, a normal distribution is secured. (See Fig. 3.) Supplementary to this assumption it is implied by the procedure of this group of test-makers that this assumption holds for the particular group of pupils to whom the exercises were given in the course of the construction of the test. The second assumption is that the variability of this normal distribution is the same for successive school grades. Both of these assumptions appear to be approximately in agreement with available data. Furthermore, they do not appear to be inconsistent with *a priori* deductions. It is true that some have contended that pupils found in the upper grades and in the high school have been selected from a total population in such a way that a skewed distribution would be obtained. This contention must be supported by additional proof before it can be accepted. In any case, it does not appear that much would be gained by assuming a skewed distribution instead of a normal distribution.

An additional assumption appears to be implied by this method of test construction. It appears to be assumed that the giving of the exercises in the preliminary list without any time limit provides adequate opportunity for pupils to demonstrate their ability.¹ This means that in judging

¹ Occasionally instead of allowing unlimited time for the entire list of exercises submitted to pupils, they are given a definite time allowance for each exercise. This time allowance is intended to be generous and to give a pupil the opportunity to do the exercise in case he has the ability.

the abilities of pupils it is sufficient to recognize only the quality of their work, and that their rate of work is not significant. As we have pointed out in other connections, the rate of work is significant in the case of some abilities. In the case of others it is relatively unimportant. This assumption is not violated in such cases.

Selection of exercises for a scaled test. In the construction of a scaled test there are three rather distinct steps. First, it is necessary to compile a preliminary list of exercises which will include a wide range of difficulty, and which will be representative of the field of the proposed test. An effort should be made to have all levels of difficulty represented, particularly the levels at the extremes of the scale. Next, this preliminary list is given to a large number of pupils in the grades for which the final test is intended. On the basis of results secured a tentative list of exercises for the final test is selected.

In making this selection, two criteria are observed. First, only those exercises are chosen which were done correctly by a gradually increasing per cent of the pupils as one proceeds from the lower to the higher grades. If an exercise is done correctly by a higher per cent of the pupils in the lower grades than in the higher grades, it is rejected as being unsuitable for testing purposes. The other criterion requires that the exercises selected be equally spaced on the scale of difficulty. It frequently happens that gaps are found so that the complete realization of this requirement is not possible. In such cases it is necessary to interpose exercises in the gaps found to exist in the preliminary list. These exercises, of course, must be given to the same group of pupils or to a similar group in order to ascertain the per cent of correct responses that may be expected. In extreme cases it will be necessary to revise the preliminary list and give the entire set of exercises to a new group of pupils.

The third step involves the evaluation of the exercises of this tentative list in terms of their difficulty, and the arrangement of them in the form of a scale. The evaluation of the exercises involves the translation of the per cents of pupils doing them correctly into difficulty values, the determination of the inter-grade interval, and the location of the zero point.

Relation between per cent of correct responses and difficulty. Ability and difficulty are correlative. The doing of an exercise having great difficulty demands the possession of a high degree of ability. The pupil who is just able to do

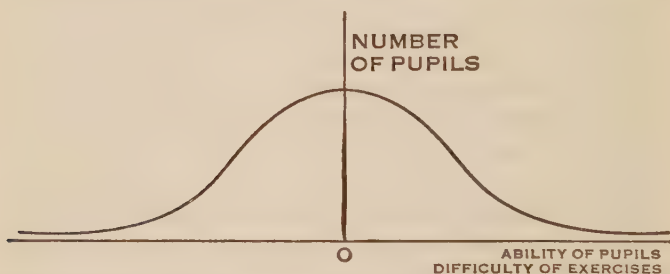


FIG. 3. A GRAPHICAL REPRESENTATION OF A NORMAL DISTRIBUTION SHOWING RELATION BETWEEN NUMBER OF PUPILS AND ABILITY

an easy exercise gives evidence of possessing only a small degree of ability. In representing the relation between per cent of correct responses and difficulty we usually represent the degrees of ability on a horizontal line. Until a more suitable zero point is established, it is customary to choose the average as zero. A point to the right of this zero represents a degree of ability greater than the average of the group of pupils being considered. A point to the left represents a less degree of ability. Vertical distances represent the number of pupils who possess exactly the degrees of ability represented on the base line. Our assumption concerning the distribution of ability in an unselected

group gives a symmetrical curve which is called the normal curve. (See Fig. 3.)

In describing a position on the base line of this normal curve a measure of the variability or spread of the curve is used. Both the median deviation (*P.E.*) and the standard deviation (σ) have been used for this purpose. If we take the degree of ability represented by $+1.00$ *P.E.* — i.e., the degree of ability represented by a position 1.00 *P.E.* to the right of our assumed zero — we find that it is possessed or exceeded by 25 per cent of the pupils. Similarly, if we take the degree of ability represented by -1.00 *P.E.* it is possessed or exceeded by 75 per cent of the pupils.

The normal curve defines a fixed relation between the two variables, ability and number of pupils. Hence, it is possible to calculate the per cent of pupils who possess or exceed any given degree of ability. The per cent of pupils who do an exercise correctly is simply the per cent of pupils who possess or exceed the degree of ability required to do that exercise. Since ability and difficulty are correlative terms, we can calculate the degree of difficulty of an exercise from the per cent of correct responses secured for it. It is only necessary to determine the per cent of pupils to the right of various points on the base line of the normal distribution. Tables have been prepared which give the per cent of pupils to the right of various points along the base line of the normal curve. In Table I we give the distances from the midpoint of the normal distribution corresponding to various per cents of correct responses. For example, there are 10.56 per cent of all pupils to the right of $+1.25\sigma$. The per cent to the right of -4.00 *P.E.* is 99.65. Also, by means of this table, the difficulty of any exercise can be determined with reference to the midpoint of the normal distribution when we know the per cent of correct responses received for it. It is only necessary to find, in the column of "per cent correct,"

TABLE I

RELATION BETWEEN PER CENT OF CORRECT RESPONSES
AND DIFFICULTY OF EXERCISES

Per cent correct	Difficulty		Per cent correct	Difficulty	
	S.D. (σ)	P.E.		S.D. (σ)	P.E.
99.999971	-5.00	-7.41	99.29	-2.45	-3.63
99.999963	-4.95	-7.34	99.18	-2.40	-3.56
99.999952	-4.90	-7.26	99.06	-2.35	-3.48
99.999938	-4.85	-7.19	98.93	-2.30	-3.41
99.99992	-4.80	-7.12	98.78	-2.25	-3.34
99.99990	-4.75	-7.04	98.61	-2.20	-3.26
99.99987	-4.70	-6.97	98.42	-2.15	-3.19
99.99983	-4.65	-6.89	98.21	-2.10	-3.11
99.99979	-4.60	-6.82	97.98	-2.05	-3.04
99.99973	-4.55	-6.75	97.72	-2.00	-2.97
99.99966	-4.50	-6.67	97.44	-1.95	-2.89
99.99957	-4.45	-6.60	97.13	-1.90	-2.81
99.99946	-4.40	-6.52	96.78	-1.85	-2.74
99.99932	-4.35	-6.45	96.41	-1.80	-2.66
99.99915	-4.30	-6.37	95.99	-1.75	-2.59
99.9989	-4.25	-6.30	95.54	-1.70	-2.52
99.9987	-4.20	-6.23	95.05	-1.65	-2.45
99.9983	-4.15	-6.15	94.52	-1.60	-2.37
99.9979	-4.10	-6.08	93.94	-1.55	-2.30
99.9974	-4.05	-6.00	93.32	-1.50	-2.22
99.9968	-4.00	-5.93	92.65	-1.45	-2.15
99.9961	-3.95	-5.86	91.92	-1.40	-2.08
99.9952	-3.90	-5.78	91.15	-1.35	-2.00
99.9941	-3.85	-5.71	90.32	-1.30	-1.93
99.9928	-3.80	-5.63	89.44	-1.25	-1.85
99.9912	-3.75	-5.56	88.49	-1.20	-1.78
99.989	-3.70	-5.49	87.49	-1.15	-1.70
99.987	-3.65	-5.41	86.43	-1.10	-1.63
99.984	-3.60	-5.34	85.31	-1.05	-1.56
99.981	-3.55	-5.26	84.13	-1.00	-1.48
99.977	-3.50	-5.19	82.89	-.95	-1.41
99.972	-3.45	-5.11	81.59	-.90	-1.33
99.966	-3.40	-5.04	80.23	-.85	-1.26
99.960	-3.35	-4.97	78.81	-.80	-1.19
99.952	-3.30	-4.89	77.34	-.75	-1.11
99.942	-3.25	-4.82	75.80	-.70	-1.04
99.931	-3.20	-4.74	74.22	-.65	-.96
99.918	-3.15	-4.67	72.57	-.60	-.89
99.903	-3.10	-4.60	70.88	-.55	-.82
99.886	-3.05	-4.52	69.15	-.50	-.74
99.865	-3.00	-4.45	67.36	-.45	-.67
99.84	-2.95	-4.37	65.54	-.40	-.59
99.81	-2.90	-4.30	63.68	-.35	-.52
99.78	-2.85	-4.23	61.79	-.30	-.44
99.74	-2.80	-4.15	59.87	-.25	-.37
99.70	-2.75	-4.08	57.93	-.20	-.30
99.65	-2.70	-4.00	55.96	-.15	-.22
99.60	-2.65	-3.93	53.98	-.10	-.15
99.53	-2.60	-3.85	51.99	-.05	-.07
99.46	-2.55	-3.78	50.00	+ .00	+ .00
99.38	-2.50	-3.70	48.01	+ .05	+ .07

TABLE I (continued)

RELATION BETWEEN PER CENT OF CORRECT RESPONSES
AND DIFFICULTY OF EXERCISES

Per cent correct	Difficulty		Per cent correct	Difficulty	
	S.D. (σ)	P.E.		D.S. (σ)	P.E.
46.02	+ .10	+ .15	0.47	+2.60	+3.85
44.04	+ .15	.22	0.40	+2.65	+3.93
42.07	+ .20	+ .30	0.35	+2.70	+4.00
40.13	+ .25	+ .37	0.30	+2.75	+4.08
38.21	+ .30	+ .44	0.26	+2.80	+4.15
36.32	+ .35	+ .52	0.22	+2.85	+4.23
34.46	+ .40	+ .59	0.19	+2.90	+4.30
32.64	+ .45	+ .67	0.16	+2.95	+4.37
30.85	+ .50	+ .74	0.13	+3.00	+4.45
29.12	+ .55	+ .82	0.11	+3.05	+4.52
27.43	+ .60	+ .89	0.097	+3.10	+4.60
25.78	+ .65	.96	0.082	+3.15	+4.67
24.20	+ .70	+1.04	0.069	+3.20	+4.74
22.66	+ .75	+1.11	0.058	+3.25	+4.82
21.19	+ .80	+1.19	0.048	+3.30	+4.89
19.77	+ .85	+1.26	0.040	+3.35	+4.97
18.41	+ .90	1.33	0.034	+3.40	+5.04
17.11	+ .95	+1.41	0.028	+3.45	+5.11
15.87	+1.00	+1.48	0.023	+3.50	+5.19
14.69	+1.05	+1.56	0.019	+3.55	+5.26
13.57	+1.10	+1.63	0.016	+3.60	+5.34
12.51	+1.15	1.70	0.013	+3.65	+5.41
11.51	+1.20	1.78	0.011	+3.70	+5.49
10.56	+1.25	+1.85	0.009	+3.75	+5.56
9.68	+1.30	+1.93	0.007	+3.80	+5.63
8.85	+1.35	+2.00	0.0059	+3.85	+5.71
8.08	+1.40	+2.08	0.0048	+3.90	+5.78
7.35	+1.45	+2.15	0.0039	+3.95	+5.86
6.68	+1.50	+2.22	0.0032	+4.00	+5.93
6.06	+1.55	+2.30	0.0026	+4.05	+6.00
5.48	+1.60	+2.37	0.0021	+4.10	+6.08
4.95	+1.65	+2.45	0.0017	+4.15	+6.15
4.46	+1.70	+2.52	0.0013	+4.20	+6.23
4.01	+1.75	+2.59	0.0011	+4.25	+6.30
3.59	+1.80	+2.66	0.0009	+4.30	+6.37
3.22	+1.85	+2.74	0.0007	+4.35	+6.45
2.87	+1.90	+2.81	0.0005	+4.40	+6.52
2.50	+1.95	+2.89	0.00043	+4.45	+6.60
2.28	+2.00	+2.97	0.00034	+4.50	+6.67
2.02	+2.05	+3.04	0.00027	+4.55	+6.75
1.79	+2.10	+3.11	0.00021	+4.60	+6.82
1.58	+2.15	+3.19	0.00017	+4.65	+6.89
1.39	+2.20	+3.26	0.00013	+4.70	+6.97
1.22	+2.25	+3.34	0.00010	+4.75	+7.04
1.07	+2.30	+3.41	0.00008	+4.80	+7.12
0.94	+2.35	+3.48	0.000062	+4.85	+7.19
0.82	+2.40	+3.56	0.000048	+4.90	+7.26
0.71	+2.45	+3.63	0.000037	+4.95	+7.34
0.62	+2.50	+3.70	0.000029	+5.00	+7.41
0.54	+2.55	+3.78			

the given per cent of correct responses and read the corresponding σ or *P.E.* value.¹ Most test-workers have used *P.E.* as the unit, although there is no reason why it should be preferred. The two units sustain a constant relation to each other: 1.00 *P.E.* equals .6745 σ .

Finding the inter-grade interval. The preliminary list of exercises should be given to groups of pupils in all grades in which the final test is to be used. Makers of this type of measuring instrument have generally calculated the difficulty of the exercises separately for each grade, and then combined the different results secured. The average ability of the different grade groups is not the same. Generally, there is a gradual increase in the average ability from a lower to a higher grade. It has been assumed that the variabilities of the distributions of the different grades were equal, which means that the difficulty values determined from different grade groups are expressed in terms of a common unit. However, the difficulty value of an exercise in the fifth grade cannot be combined with the difficulty value of an exercise in the seventh grade until these values are expressed with reference to the same zero point. To do this requires the determination of the inter-grade interval, or the distance between the assumed zero points of the distributions of ability in the successive school grades.

Three methods have been proposed for making this determination:

1. Exercise method.
2. Quartile method.
3. Distribution method.

¹ Several different forms of tables have been used in expressing the relation between the per cent of correct responses and the difficulty of exercises. In order to economize space, most tables have not been extended beyond 50 per cent; but the normal distribution which we get in such a table can easily be adapted so as to yield determinations for per cents above 50. Some tables have been expressed in terms of per cents of incorrect responses rather than in terms of correct responses.

1. *Exercise method.* Since each exercise is given to pupils in different school grades, the difference in its value for two successive grades, when both values are expressed as a distance from the midpoint of the distributions of the respective grades, is easily determined. If we assume that these values expressed with reference to the absolute zero point would be identical, the difference between them is the difference between the assumed zero points, or the midpoints of the distributions of these two grades. The method may be illustrated by example 10 of Woody's Addition Scale. It was found to be 1.629 *P.E.* below the midpoint of the third-grade distribution, and 2.397 *P.E.* below the midpoint of the fourth-grade distribution. The difference (inter-grade interval) is .768 *P.E.* In this way the inter-grade interval between two successive grades may be found for each exercise that was attempted by pupils in both grades.

The determinations of the inter-grade interval secured in this way will not be equal. In general, the more difficult the exercise is for the pupils of the grades concerned the greater the inter-grade interval will be. For this reason, it has been proposed that, instead of taking the average of all the determinations of the inter-grade interval, the determinations should be weighted so as to give the greater weight to those determinations secured from moderately easy or moderately difficult exercises.

2. *Quartile method.* By this method the performance of each pupil is described in terms of the number of exercises done correctly. The scores thus obtained tend to form a normal distribution for each grade. The median and the quartile deviations are calculated for each distribution. The difference between the medians of two successive grades is divided by the average of the corresponding quartile deviations. For example, if the median for the second grade is 6.8, and the one for the third grade is 14.5, the difference is

7.7. If the corresponding quartile deviations are 2.7 and 3.0, the average is 2.85. Using this as a divisor into 7.7, the inter-grade interval is found to be 2.7 quartiles or *P.E.*

3. *Distribution method.* The calculation of the inter-grade interval by this method is based upon the fact that the distribution for one grade overlaps those of the adjoining grades. If the number of examples done correctly is used as the pupil's score, there will be some pupils in the third grade who have higher scores than the median score of the fourth grade, and others who are below the median score of the second grade. The per cent of pupils who have scores between the median for their grade and the median for another grade can be used as the basis for calculating the inter-grade interval. Such a per cent is merely the per cent of pupils between the midpoint of a normal distribution and a point to the right or left whose distance is to be determined. If the point is to the left, i.e., if it is the inter-grade interval between a lower grade, 50 should be added to the per cent of pupils and then the inter-grade interval may be read directly from Table I. If the distance is to the right of the zero point, i.e., if it is the inter-grade interval between a higher grade, the per cent of pupils must be subtracted from 50 before Table I can be used.

In applying the distribution method it is not necessary that adjacent grades be used. Any two grades between which there is overlapping may be used. The interval obtained when the two grades are not adjacent is not the inter-grade interval, but rather the sum of the inter-grade intervals between the two grades used. For example, if the third and fifth grades have been used, the interval will be from the median of the third grade to the median of the fifth grade. Knowing this distance, and having calculated the inter-grade interval from the third to the fourth grade, we can obtain a determination of the inter-grade interval from the

fourth to the fifth grade. Such secondary determinations are not as significant as the primary determinations and should be given less weight in calculating the average inter-grade interval.

Combining the results of the three methods. Some test-makers have used all three methods and have calculated a weighted average of the three determinations. Others have used only one method. It appears that the exercise method is considered to yield the most significant results and, hence, the determination by this method has been given more weight than the other determinations in calculating the average.

Locating the zero point. In all of the preceding work the difficulty values of the exercises have been expressed as distances from the midpoint of the grade distributions. With the inter-grade intervals known, it is possible to express all distances from a common zero point, such as the median of the third-grade distribution.¹ If this were done some distances would still be negative, and such a scale would not be as convenient to use as one which involved only positive quantities. Obviously, negative quantities can be eliminated by choosing the zero point sufficiently far to the left and expressing all values with reference to it. There is, however, one disadvantage in locating the zero of a scale in this way. We are unable to say that an example that has a scale value of 5 is just half as difficult as one that has a scale value of 10, because, if we had located the arbitrary zero 4 units farther to the left, the scale values of these two exercises would have been 9 and 14, respectively. Had the zero point been located 4 units to the right, their scale values would have been 1 and 6, respectively. Before a scale value of 10 can be interpreted to mean twice the difficulty of a scale value of 5, both must be expressed with reference to an absolute zero

¹ This assumes that the exercises were given to pupils of the third grade.

point. Such a point would represent not any of the ability in question.

The methods used in locating the zero point are more or less arbitrary, and due caution must be exercised in interpreting the zero points chosen as meaning not any of the ability in question. A common method is to note the per cent of pupils failing to do a single exercise correctly in the lowest grade in which the exercise has been given. These pupils are considered not to possess any of the ability being measured. The remaining per cent of the pupils (100 minus the per cent doing no exercise correctly) possess some ability; i.e., they are above the absolute zero point. By referring to Table I the distance of this zero point from the midpoint of the distribution may be ascertained. The same method may be applied to the distribution for any grade which includes zero scores. The zero point determined in this way approaches a true absolute zero point only when exercises calling for the least definable amount of ability are included in the list submitted to the pupils.

Woody has also calculated the zero point by dividing the median number of examples done correctly by the quartile deviation of the distribution. This method, however, appears to be less satisfactory than the one described above.

Determining the final scale values of the exercises. Knowing the inter-grade interval and the location of the zero point, it is now possible to calculate the final scale value for each exercise. This value will denote its distance from the zero point chosen. It is calculated from the determinations in the several grades. These, when corrected for the inter-grade intervals, will differ rather widely in some cases. This is most likely to happen when the exercise is either very easy or very difficult for the pupils of a given grade. Such determinations should be given less weight than those secured when the exercise is done correctly by from 25

to 75 per cent of the pupils. Woody gives a value double weight if the exercise is less than 1 *P.E.* distant from the median achievement of a grade distribution, single weight if it is more than 1 *P.E.*, but less than 3 *P.E.* distant from the median achievement, and he does not consider it at all if it is more than 3 *P.E.* distant from the median. The weighted average which he obtained from this method is taken as the general or final value of the exercise.

After the values of all of the exercises have been determined, a group of exercises for the final scale should be selected. If possible those exercises should be chosen which are equally spaced on the scale of difficulty. Unless a large number of exercises have been evaluated, it may happen that no such group can be selected. If it is deemed imperative to have a group of exercises that are equally spaced on the scale of difficulty, it will be necessary to prepare another list of exercises and repeat the procedure just described.

QUESTIONS AND TOPICS FOR DISCUSSION

1. What requirements should govern the construction of educational tests?
2. What requirements of test construction appear to have been recognized by the makers of the following tests:
 - a. Woody's Arithmetic Exercises?
 - b. Buckingham's Spelling Scale?
 - c. Ayres's Scale for Measurement of Ability in Spelling?
 - d. Burgess's Picture Supplement Scale for Measuring Ability in Silent Reading.
 - e. Courtis's Standard Research Tests in Arithmetic, Series B?
3. Did the authors of the above tests overlook any essential requirements? If so what ones?
4. One of the requirements to be observed in test construction is that all pupils be given an opportunity to demonstrate their abilities in the field of the test. Is this requirement adequately provided for in the following tests:
 - a. Courtis's Standard Research Tests in Arithmetic, Series B?
 - b. Woody's Arithmetic Exercises?
 - c. Cleveland Survey Tests in Arithmetic?
 - d. Monroe's Standardized Silent-Reading Test?

- e. Burgess's Picture Supplement Scale for Measuring Ability in Silent Reading?
- f. Charters's Diagnostic Tests in Language and Grammar?
5. Distinguish between a "power test" and a "speed test." (See McCall, W. A., *Correlation of Some Psychological and Educational Measurements*. Teachers College Contributions to Education, No. 79, pp. 44-53.)
6. What types of tests may we recognize from the standpoint of structure?
7. Does the available evidence justify us in assuming that indirect measurement is possible? (Buckingham, B. R. "Proposed Index to Efficiency in Teaching U.S. History"; in *Journal of Educational Research*, vol. 1, p. 161 (March, 1920).

SELECTED REFERENCES

AYRES, L. P. *A Measuring Scale for Ability in Spelling*. Division of Education, Russell Sage Foundation, Bulletin No. 139, New York City, pp. 56.

BUCKINGHAM, B. R. "Principles of Scale Derivation; with Special Application to Arithmetic, Geography, History, and Grammar"; in *Proceedings of Third Indiana Conference on Educational Measurements*, 1917, pp. 49-84.

BUCKINGHAM, B. R. "Notes on the Derivation of Scales in School Subjects; with Special Application to Arithmetic"; in *Fifteenth Yearbook of the National Society for the Study of Education*, part 1, 1916, pp. 23-40.

BUCKINGHAM, B. R. *Spelling Ability, its Measurement and Distribution*, p. 80. Teachers College Contributions to Education, No. 59.

BURGESS, MAY AYRES. *The Measurement of Silent-Reading Ability*. Division of Education, Russell Sage Foundation.

CHARTERS, W. W. "Constructing a Language and Grammar Scale"; in *Journal of Educational Research*, vol. 1, pp. 4, 249-57.

COUNTS, GEORGE S. *Arithmetic Tests and Studies in the Psychology of Arithmetic*. Supplementary Educational Monographs, University of Chicago Press, 1917, p. 120, vol. 1, No. 4 (August, 1917).

COURTIS, S. A. *The Curtis Tests in Arithmetic*. Final report, Committee of School Inquiry, Board of Estimate and Apportionment, New York City, vol. 1, 1911, 1912, part 2, pp. 389-546.

KELLEY, T. L. "Simplified Method of Using Scaled Data for Purposes of Testing"; in *School and Society*, vol. iv, pp. 34-7, 71-5 (July 1-8, 1916).

LACKEY, E. E. "A Scale for Measuring the Abilities of Children in Geography"; in *Journal of Educational Psychology*, vol. ix, 1918, pp. 443-51.

MCCALL, W. A. *Correlation of Some Psychological and Educational Measurements*. Teachers College Contributions to Education, No. 79, pp. 44-53.

MINNICK, J. H. *An Investigation of Certain Abilities Fundamental to the Study of Geometry*. Philadelphia: University of Pennsylvania, 1918.

MONROE, W. S. "An Analytical and Experimental Study of Woody's Arithmetic Scales"; in *School and Society*, vol. vi, pp. 412-20 (October, 1917).

MONROE, W. S. "A Series of Diagnostic Tests in Arithmetic"; in *Educational School Journal*, vol. XIX, 1918-19, pp. 585-607.

MONROE, W. S. Report of Division of Education Tests for 1919-20. Bulletin No. 5, Bureau of Educational Research, University of Illinois, pp. 36-47.

RUGG, H. O., and CLARK, J. R. *Scientific Method in the Reconstruction of Ninth-Grade Mathematics*. Supplementary Educational Monographs, University of Chicago Press, 1918, vol. II, No. 1 (April, 1918).

STARCK, D. "A Scale for Measuring Ability in Arithmetic"; in *Journal of Educational Psychology*, vol. VII, pp. 213-22 (April, 1916).

THEISEN, W. W. and FLEMMING, CECILE W. "The Diagnostic Value of the Woody Arithmetic Scales"; in *Journal of Educational Psychology*, vol. IX, 1918, pp. 475-88.

THORNDIKE, E. L. "The Measurement of Educational Products"; in *School Review*, vol. XX, pp. 289-99 (May, 1912).

THORNDIKE, E. L. "An Improved Scale for Measuring Ability in Reading"; in *Teachers College Record*, vol. XVI (November, 1915); and vol. XVII, pp. 40-67 (January, 1916).

THORNDIKE, E. L. "Measurement of Achievement in Reading; Word Knowledge"; in *Teachers College Record*, vol. XVII, pp. 430-54 (November, 1916).

THORNDIKE, E. L. "Tests of Intelligence; Reliability, Significance, Susceptibility to Special Training, and Adaptation to the General Nature of the Task"; in *School and Society*, vol. IX, pp. 189-95 (February 15, 1919).

WOODY, CLIFFORD. *Measurements of Some Achievements in Arithmetic*. Teachers College Contributions to Education, No. 80.

CHAPTER V

DESCRIPTION OF THE PERFORMANCES OF PUPILS

A quantitative scale is required for description. A scale is necessary for quantitative description. It is a part of every measuring instrument. In some it appears in very obvious form. In others it is implied in the structure of the test, the rules for scoring, and the method of computing scores. The complete description of a pupil's performance requires a quantitative statement of its quality and of the rate at which it was produced, together with the specifications of the exercises in response to which the performance was given. These specifications are expressed in various forms. For the Courtis Standard Research Tests in Arithmetic, Series B, a statement of the structure of the exercises is sufficient. For example, in the case of the addition test it is sufficient to say that each exercise consists of three columns of figures, with nine figures to the column. In other fields the specifications cannot be given in such a simple form. One of the forms in which the specifications are sometimes given is in terms of the difficulty of the exercises of the test.

The method of describing a performance varies for each of the three characteristics, and also for different types of performances. Some tests provide a separate description for each of the three characteristics. Other measuring instruments combine two and, in a few cases, all three in a single score. It will, therefore, be necessary to consider separately the description of each of the three characteristics with reference to the different types of performance. After this we shall note some of the methods of description which combine two or more of the characteristics in a single score.

Rate of uniform performances. In order that a measure of the rate at which a performance has been produced may be secured, two requirements are necessary. First, the performance, or some representative portion of it, must be timed; and, second, the timed portion of the performance must be of such a nature that it can be divided into equal work units,¹ or into exercises consisting of known multiples of a convenient work unit. For group testing the plan generally used, and the most convenient one, is to obtain only the performance that pupils are able to produce within a suitable time interval. Another plan which is sometimes used is to have each pupil indicate the portion of the performance which he completes within a given time interval. This second plan is not very satisfactory. In individual testing it is possible to time each pupil on either the entire performance or a portion of it, as may be desired. The number of unit performances given within the time interval constitutes a quantitative statement of the pupil's rate of work. For convenience the rate is frequently expressed in terms of the number of units of work done within a unit of time, such as a minute or second.

The structure of the test usually determines what the work unit shall be. In some subject-matter fields, such as the operations of arithmetic, it is possible to construct the exercises so that they form approximately equal units of work. The structure of a number of tests is such that it is not convenient to divide the pupil's performance into work units which are approximately equal. For example, in the case of silent reading, the most convenient work unit to use, and the one which is generally used, is the word. It is obvious that one word is not necessarily equal to another word. However, the variations of particular words from an "aver-

¹ "Equal work units" may be defined as divisions of a test which in general are done in equal time intervals and with a constant quality.

age word" are chance variations, and when the performance is sufficiently long the word does constitute a satisfactory unit for measuring the rate of work. A similar situation exists in the case of handwriting. The letter is usually taken for the unit of work, but one letter is not equal to another letter.

Rate of non-uniform performances. In irregular tests the exercises do not represent equal units of work. However, when the irregularities are not large and are distributed in a random manner, and the number of units is sufficiently large, the error introduced by considering the exercises to represent equal work is slight. Thus the rate may be described in terms of the number of exercises attempted, as in the case of uniform performances.

In the case of scaled performances, or performances with marked irregularities, the rate is probably not constant throughout the performance. One would expect it to be more rapid on the easier exercises, and less rapid on the more difficult ones. Thus the exercises do not represent equal work units, and it is difficult, if not impossible, to divide the performance into equal work units. For this reason the use of scaled performance tests or of tests with marked irregularities should be avoided, if possible, when the rate of work is important. Several plans for weighting the exercises have been proposed, but they are unsatisfactory for securing a true measure of the rate of the performance. Certain systems of weighting which have been used will be described in the discussion of the description of the quality of non-uniform performances.

Two measures of quality. In some subject-matter fields, such as arithmetic or spelling, we are accustomed to judge the quality of a performance as either right or wrong. Imperfections are definite errors. In such subject-matter fields quality becomes accuracy. In other subject-matter fields

imperfections are not errors in the sense that eleven would be an erroneous sum for $7 + 5$. In English composition, for example, there are some imperfections which are errors, but a particular composition may in many ways fall short of perfection, not because of errors, but because certain qualities are not present in the fullest possible degree. Frequently these qualities are subtle. What is true of English composition and other similar performances is true to some extent of silent reading, oral reading, and the answers to certain kinds of questions, but in these subjects exercises have been constructed and rules for scoring devised so that the description of performances is essentially in terms of accuracy.

When a pupil's performance on an exercise cannot be classified as right or wrong, or at least as right, partially right, or wrong, quality becomes a matter of merit rather than of accuracy. It will therefore be necessary to make certain differentiations in our discussion of the description of the quality of performances. We shall consider first quality when it is defined as accuracy, and second when it is defined as merit. The description of the latter will be discussed in the next chapter.

Accuracy of uniform performances. The usual form of describing the accuracy of a uniform performance requires that it be divided into equal units of work. The accuracy of the performance is then expressed as the per cent of these units in which a certain standard of accuracy has been attained. The expression of the accuracy as a per cent is not essential except for purposes of comparison when all pupils do not attempt the same number of exercises. If sufficient time is allowed, so that all pupils have had the opportunity to attempt all of the exercises which they are able to do correctly, we may describe the quality of their performances in terms of the number of units done correctly. However, the measure is probably more easily understood when ex-

pressed as a per cent. It should be noted that when expressed as a per cent the accuracy is an index of the average quality of the performance. It must not be interpreted to mean that throughout the entire performance the pupil was accurate only to the extent indicated by the accuracy score. As a matter of fact his performance was perfect for some units and for other units was incorrect.

The scale of accuracy. The scale of accuracy when expressed as a per cent is from 0 to 100. This, however, does not mean that a pupil's per cent of accuracy may always be any number between 0 and 100. If the performance consists of 10 units and partial credits are not given, the possible measures of accuracy are 0, 10, 20, 30, etc. If the performance consists of 20 units the possible measures of accuracy are 0, 5, 10, 15, 20, 25, etc.

Standards of accuracy. An accuracy score of zero per cent does not necessarily mean that the pupil did absolutely nothing correctly. For example, in the addition test of the Courtis Standard Research Tests, Series B, an answer is counted as entirely wrong if one figure is incorrect. If a pupil has no answer entirely correct, his accuracy score is zero, even though all figures except one may be correct in every example. For the same reason, an accuracy score of 40 per cent does not mean that in the remaining 60 per cent of the units of work the pupil did nothing correctly. It simply means that in 60 per cent of the units of work the pupil failed to attain the required standard of accuracy. Thus the meaning of an accuracy score depends upon the standard of accuracy which has been set for judging the units of the performance.

The units of most performances may be considered to consist of elements or of opportunities to make errors. For example, an exercise in the addition of integers consisting of three columns of figures of nine figures each involves the

addition of twenty-seven numbers, plus two numbers to be carried. Each figure must be read and the answer must be written. Such a unit may be analyzed in several ways, but it is obvious that any form of analysis will give a large number of opportunities for making an error. The standard of accuracy depends upon the length of the unit exercise, i.e., upon the number of opportunities for error, and upon the penalty to be charged against the pupil for making an error.

For purposes of illustration, take two exercises in the addition of integers:

$$\begin{array}{r} \text{(A) } 6 \\ 3 \\ \hline \end{array} \qquad \begin{array}{r} \text{(B) } 4 \ 5 \ 7 \ 2 \ 4 \ 6 \ 3 \ 5 \\ 2 \ 3 \ 1 \ 7 \ 5 \ 2 \ 4 \ 2 \\ \hline \end{array}$$

Both of these examples consist of addition combinations whose sums are less than ten. Hence no carrying is involved in either example. Assuming that for the same pupil the chance of an occurrence of error is the same for all of these combinations — i.e., that he knows them equally well — the opportunities for error are eight times greater in example B than in example A. Consequently, according to the usual rule for scoring such examples, success on example A does not mean the same as on example B. Success on example B requires a higher quality of work. By increasing the number of opportunities for error it would be possible to set an example containing a sufficient number of elements so that it would be done correctly by practically no pupil of the elementary school. This would probably be true if the number of elements were increased to 1000. Although the pupil might “know” all of these addition combinations, which are the simpler addition combinations, human fallibility would eventually enter in to produce an error.

The standard of accuracy also depends upon the penalty which is levied against a performance for an error. Con-

sider two tests upon the multiplication combinations. Each test consists of 100 combinations. In the first test each combination is considered a unit exercise. A pupil who makes twelve errors will be penalized twelve units or combinations, and the accuracy of his performance will be 88 per cent of perfection, defining perfection to mean the writing of 100 combinations without error. The second test is divided into unit exercises of five combinations each. The test will consist of twenty of these units. Suppose a pupil in doing these twenty exercises (100 combinations) writes twelve combinations incorrectly. These errors, if we assume a random distribution of the combinations within the test, are probably distributed rather uniformly over these 100 combinations. The maximum number of wrong exercises would be twelve and the minimum three. The true number is probably between these two extremes, although nearer twelve than three.

The usual plan of scoring is to penalize a pupil for an entire unit if that unit contains one error. The same penalty is levied against his performance if the unit contains more than one error. According to this plan, if all of the errors were concentrated in three units, the per cent of accuracy would be 85. If the twelve errors were distributed so that there was only one in a unit exercise, the per cent of accuracy would be 40. These are the extreme possibilities. If the errors were uniformly distributed over the performance, the per cent of accuracy would fall between these extremes, but in no case would it be equal to that on the other test. A pupil's accuracy score on this second test would differ from his accuracy score on the first test, not by reason of the actual number of errors made, but by reason of the penalty which was levied for an error.

Arbitrary standards of accuracy. In arithmetic or spelling the standard of perfect accuracy is usually adopted. It

has the approval of social usage. But other standards could be set up. For example, in arithmetic an answer could be counted correct if only one figure was wrong, or the spelling of a word could be considered correct if the sound of the letters used agreed with the pronunciation of the word. It would be absurd to set up such standards in the operations of arithmetic or in spelling except for a special investigation, but for tests in other fields similar standards are frequently established. In oral reading Gray has set a standard which involves giving full credit when only one error is made. In scoring answers to questions that require pupils to formulate answers in their own words, considerable variation is generally allowed in the answers that are accepted as correct. Thorndike does this in his Scale Alpha 2 for the Understanding of Sentences. Partial credit is sometimes given for answers which are judged to be partially right.

Two types of uniform performances. In such subject-matter fields as the operations of arithmetic, where the exercises of a test may be constructed so that they contain the same number of elements, a uniform performance is generally one in which the exercises consist of the same number of elements and consequently imply the same standards of accuracy. In other types of subject-matter fields different exercises may consist of different numbers of elements, although they are equally difficult. As a result the standards of accuracy are likely to be different. The per cent of correct responses obtained for an exercise depends upon two things: the number of opportunities for error, and the chance of occurrence of error in an element. For a given per cent of correct responses these two factors vary inversely. The greater the number of elements the less the chance of the occurrence of error in an element, and conversely. We may, therefore, have a uniform performance in which the units contain the same number of opportunities and the chance of

the occurrence of error is the same for all units. We may, however, have a uniform performance in which the number of elements per unit is different, and the chance of occurrence of error in an element likewise different. For practical purposes we do not need to distinguish between these two types of uniformity.

Equivalent standards of accuracy necessary for comparable measures. In our discussion of standards of accuracy we showed that for the same performance different descriptions would be obtained, depending upon the standard of accuracy used. It therefore follows that measures of accuracy are not comparable unless the standards of accuracy on which they are based are equivalent. The standards of accuracy will be equivalent when the unit exercises upon which both are based contain the same number of elements, and the penalty levied against the performance for the occurrence of error is the same. It is only when these conditions are fulfilled that comparisons of accuracy scores have significance. For example, on the basis of the four tests of the Courtis Standard Research Tests, Series B, as they are usually scored, it is impossible to make any inferences concerning the relative accuracy of pupils in the different tests. It has been inferred from the scores resulting from the use of this series of tests that addition was a more difficult process than subtraction, because the average per cent of accuracy is greater for subtraction than for addition. Even a casual examination of these two tests will reveal that the standards of accuracy by which the performances of pupils are judged on these two tests are not equivalent. The penalty for error is much more severe in the case of addition.

Another condition must be fulfilled if small differences between measures of accuracy are to be significant. If a performance consists of three units, the per cent of accuracy is either 0, $33\frac{1}{3}$, $66\frac{2}{3}$, or 100. If a performance consists

of 10 units the possible per cents of accuracy are 0, 10, 20, etc. If a performance consists of 50 units the possible per cents of accuracy are 0, 2, 4, 6, 8, etc. In comparing the per cent of accuracy on a performance of 50 units with the per cent of accuracy of a performance consisting of 3 units, it is clear that no significance can be attached to small differences between the two measures. A similar statement may be made with reference to the comparison of individual scores with norms.

In order that small differences between the measures of accuracy may be significant it is necessary that both of the performances consist of a large number of units. This requirement is generally violated for some pupils by allowing the same time interval for all pupils when a timed performance is secured. Since some pupils work much more rapidly than others, the performances of the slower pupils will contain only a few units. In order to secure performances which are approximately equal in length, it will be necessary to modify the timed performance tests which are now commonly used. One plan that has been used by certain test-makers is to have the pupil mark the place reached at specified intervals. Courtis has done this in his Silent Reading Test No. 2. This plan is only partially satisfactory.

Accuracy of non-uniform performances. In describing the accuracy of a non-uniform performance, it appears reasonable that more credit should be given for doing a very difficult exercise correctly than for giving the correct response to a very easy one. Several attempts have been made to devise a method of weighting the exercises which would compensate for the differences in their size and difficulty. In most cases exercises of different levels of difficulty imply different standards of accuracy. When the answer to a long and intricate example in arithmetic is called wrong because a single figure is incorrect, the standard of

accuracy is distinctly different from that which is used in judging the correctness of a simple example. This makes it difficult to devise a procedure which will yield a valid description of the average accuracy of the performance and which may be easily interpreted. Several of the proposed plans are described below. After weights have been assigned to the exercises according to a given plan, the sum of the weights, or values, of the exercises done correctly is usually taken as the pupil's score. In case the test has been timed so that different pupils have attempted different numbers of exercises, such a score is a combination of rate and quality.

Plans for weighting exercises of non-uniform performances. 1. Fordyce, in constructing his silent-reading test, assigned values to the exercises that are proportional to the square root of the per cent of incorrect responses. He states that this is in accord with the law of least squares. This is an effort to give more credit for doing a difficult exercise than for doing an easy one.

2. One writer has proposed that we assign values proportional to the number of incorrect responses.¹

3. Weights have been assigned to exercises proportional to the degree of difficulty expressed in terms of the mean deviation (*P.E.*), or standard deviation with the zero point fixed arbitrarily.²

4. In the derivation of the Kansas Silent Reading Tests, F. J. Kelly obtained for each exercise the average number of seconds per correct answers. Some of the exercises were longer than others and hence required more time. Some were short but difficult in the sense that the per cent of cor-

¹ Brooks, S. S. "Getting Teachers to Feel the Need for Standardized Tests"; in *Journal of Educational Research*, vol. II, pp. 431 (June, 1920).

² Monroe, Walter S. *Report of Division of Educational Tests for 1919-20*, University of Illinois Bulletin, vol. XVIII, No. 31 (January 24, 1921).

rect responses was small. The length of time required to do the exercise and the degree of difficulty were combined by this method.¹

5. It has been pointed out that the allowance of more credit for difficult exercises than for easy ones does not inflict upon a pupil the proper penalty for failing on a simple exercise. It has, therefore, been suggested that the best description of a pupil's performance will be expressed in terms of a system of credits for correct responses, and penalties for incorrect responses.²

6. The regression equation is probably the most scientific method for weighting exercises. This method requires that we have an independent criterion of the ability being measured. This is difficult to secure. In addition, the use of the regression equation involves an intricate statistical procedure.³

7. The above plans of weighting, except the last, are explicitly for the purpose of adjusting the credits given to the difficulty of the various exercises. These proposals imply that the significance of a pupil's performance on an exercise is a function of the difficulty of the exercise. In defining our educational objectives we generally assume that the importance of topics is determined by their usefulness. Many very difficult exercises are unimportant because they are not useful. For this reason it has been proposed to weight exercises on the basis of their social importance. According to this proposal the doing of a very difficult exercise might add

¹ Kelly, F. J. "Kansas Silent Reading Tests"; in *Journal of Educational Psychology*, vol. vii, pp. 62-80 (February, 1916).

Monroe, Walter S. "Monroe's Silent Reading Tests"; in *Journal of Educational Psychology*, vol. ix, pp. 303-12 (June, 1918).

² Minnick, J. H. *An Investigation of Certain Abilities Fundamental to the Study of Geometry*. Philadelphia: University of Pennsylvania, 1918.

³ Kelley, T. L. *Vocational Guidance*, p. 95. Teachers College Contributions to Education, No. 71.

little to a pupil's score. This method demands a criterion of social usefulness. Because this is difficult to secure, practically no use has been made of the method.

Scores based upon unweighted exercises. Any plan of weighting adds considerably to the task of constructing the test and to the scoring of the test papers. For this reason it has been proposed that the unequal difficulties of the exercises be disregarded, and that scores be calculated just as though the exercises were equal in difficulty. Although this procedure appears to be illogical, it has been found that the correlation between weighted and unweighted scores is extremely high. Charters dropped the weights from the exercises in his language and grammar tests because he felt that the differences in the scores obtained by the two methods were not sufficient to justify the added labor involved in computing the weighted scores.

Description of a scaled performance. Since a scaled performance is not usually timed, the consideration of the description of the rate of such a performance may be omitted. The description of quality is implied in the statement of the highest level of difficulty reached with the fixed standard of accuracy. Hence the description of a scaled performance is confined to a single score which is intended to define the step of the scale which the pupil was barely able to reach with a fixed standard of accuracy.

The theory of this type of measure is that the pupil is presented with a series of exercises or groups of exercises which gradually increase in difficulty. The situation has been likened to such athletic events as the pole vault, high jump, and so forth. As the pupil advances along the scale he will finally reach the level where he does not have the ability (power) to do the exercises with the established standard of accuracy. When this ideal condition exists a pupil's ability (power) is described in terms of difficulty of the highest step

of the scale which he does satisfactorily. However, the typical pupil will not do all exercises or groups of exercises up to a certain point with a degree of accuracy up to or above the standard set, and fail to attain this standard of accuracy on all succeeding exercises. He will fail to attain the standard accuracy for certain exercises, and then attain or exceed it for more difficult ones. For a considerable range of the scale he will alternate between success and failure. These fluctuations complicate the problem of describing a pupil's performance on tests of this type. It is necessary to secure a performance on a series of levels of difficulty in the neighborhood of the theoretical level, on which a pupil just barely attains the standard of accuracy. Of course, in group testing, where a single test is applied to all pupils under the same conditions, it is necessary to have a considerable range of difficulty in order that each one may give a performance on an appropriate series of levels of difficulty.

Relation of accuracy to difficulty. The study of the performances of pupils on successive levels of difficulty has indicated that the relation between the quality or accuracy of the performance and the levels of difficulty is represented by the ogive curve. The equation of this curve is:

$$y = k \int e^{-\frac{x^2}{2\sigma^2}} dx$$

This equation will be recognized as an integral of the equation of the normal distribution:

$$\frac{dy}{dx} = K e^{-\frac{x^2}{2\sigma^2}}$$

In the above equation y stands for the quality of the pupil's performance; i.e., the per cent of exercises done correctly on a given level of difficulty. The other variable, x , stands for the difficulty of the given group of exercises. In Table I

(see page 96) we have given the values of y which correspond to values of x between -5σ and $+5\sigma$. When these values are represented graphically an ogive curve is obtained. In Fig. 4 the zero point has been chosen arbitrarily at -8 . With this modification it is a representation of Table I.

Standards of accuracy with respect to which level of difficulty is expressed. This relationship between quality

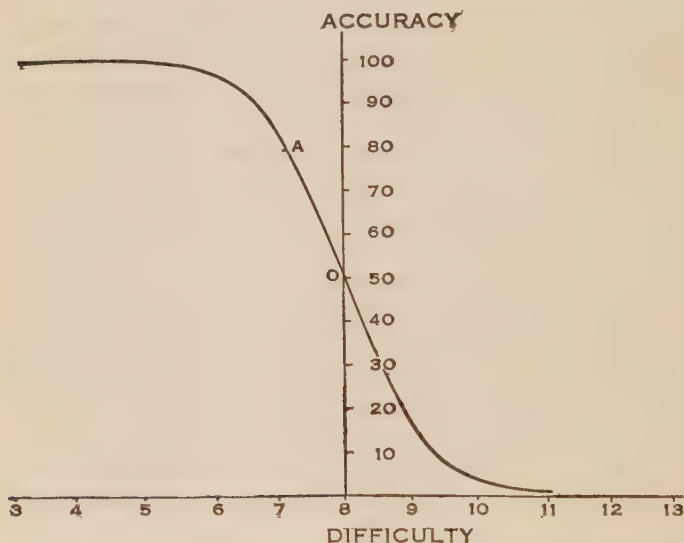


FIG. 4. THE OGIVE CURVE, SHOWING THE RELATION BETWEEN ACCURACY AND DIFFICULTY. THIS IS A GRAPHICAL REPRESENTATION OF TABLE I WITH THE ZERO POINT AT -8 .

of performance and level of difficulty appears to hold for both individual pupils and for groups of pupils. In Fig. 4 and in Table I it is clear that near 100 per cent accuracy, or near zero accuracy, the change in accuracy corresponding to a unit change in difficulty is very small. For example, in

Table I the change in accuracy between a level of difficulty of -5σ and a level of difficulty of -4σ is .003171 per cent; between -4σ and -3σ it is .1318 per cent. On the other hand, change from -1σ to 0σ is 34.13 per cent. A similar condition exists on the higher levels of difficulty. If the fixed standard of accuracy is chosen near 100 per cent, it is obviously necessary to have a very large number of exercises on each level of difficulty, or to have some means for ascertaining very precise measures of the accuracy of a pupil's work. In view of the variable character of pupil performances a high degree of precision in expressing the degree of accuracy cannot have much significance. For these reasons, it is necessary to choose a standard of accuracy for which the ratio of the rate of change of accuracy to the rate of change of difficulty is near a maximum. This ratio is at a maximum for 50 per cent accuracy. It recedes from this maximum rather slowly for small distances on either side of a standard of 50 per cent. Pedagogical reasons may be given for using a standard of accuracy higher than 50 per cent. Thus some makers of tests have chosen a standard of accuracy of 80 per cent. In one case a standard of accuracy of 75 per cent has been used. From the standpoint of measurement a standard of 50 per cent is to be preferred.

Estimating the level of difficulty corresponding to a given standard of accuracy. Because of the irregularities in the accuracy of pupil performances, and the imperfections of difficulty scales, it is generally necessary to estimate the level of difficulty on which the established standard of accuracy would be attained. This estimate is based upon the quality of the pupil performance on the levels of difficulty near this point on the scale. If we assume that for a given pupil or a given group of pupils the relation between the accuracy of their performances and the levels of difficulty on which these performances are given is that represented

by the ogive curve, it is possible to estimate, from a pupil's accuracy on known levels of difficulty, the theoretical level of difficulty on which the established standard of accuracy would be obtained.

Suppose that on the level of difficulty 7.1σ a pupil has rendered a performance which is 80 per cent accurate. The ogive curve which corresponds to the degree of ability possessed by this pupil must pass through the point whose coördinates are 7.1σ and 80. In Fig. 4 this is curve A. We wish to know the level of difficulty on which he would be most likely to render a performance which is 50 per cent accurate. If a horizontal line is drawn to represent 50 per cent of accuracy it will cut this ogive curve at O. A perpendicular from this point will cut the horizontal axis, the point which represents the level of difficulty on which a performance 50 per cent accurate may be expected. This point is 8.0. This method is entirely general. It is only necessary to determine the position of the particular ogive curve corresponding to the ability of a particular pupil. This is done by locating the point whose coördinates are the per cent of accuracy and the difficulty on a level of the test.

In calculating a pupil's score on a scaled performance test, the level of difficulty on which he will meet the standard of accuracy is estimated from his performances on several levels of difficulty. The usual procedure is to use all performances on all levels except those for which the per cent of accuracy is approximately either zero or 100. The pupil's score is the average of these several estimates. This graphical procedure is obviously too tedious for practical use. Thorndike has constructed tables for estimating the level of difficulty on which his standard of accuracy (80 per cent) is attained on his Visual Vocabulary Scales, and also on his Scale Alpha 2 for the understanding of sentences. His tables are, however, not satisfactory for computing in-

dividual scores. T. L. Kelley has prepared a more elaborate table to be used in computing such scores on Alpha 2.¹

Kelley's shorter method of estimating scores. T. L. Kelley² has proposed a "simplified method" for computing the scores of individual pupils. His procedure assumes that the differences between successive levels of difficulty are equal. If this assumption is not made the procedure is greatly complicated. According to his method, the pupil's performance upon each level of difficulty is described with reference to accuracy. If the performances on each level consist of the same number of units, this description may be given in terms of the number of units done correctly. In case a performance which ranges in quality from perfect accuracy to zero accuracy has not been obtained, it is necessary to estimate extensions of the pupil's performance to include these limits. These estimates are based upon the relation which is represented by the ogive curve. The differences between the accuracy on successive levels of difficulty are then calculated. Because pupils tend to be erratic in their performances some of these differences may be negative. These differences are next added algebraically, beginning with the ones for the lower levels of difficulty, until the sum is half of the magnitude of the scale of accuracy. For example, if the accuracy of the performance has been described in terms of per cent, half of the magnitude of the scale of accuracy would be 50. If each level of difficulty consists of 10 units, and the accuracy has been described in terms of the number of units done correctly, half of the scale would be 5. The level of difficulty reached when the partial sum of the differences is equal to the half sum is the level of difficulty

¹ Kelley, T. L. "Thorndike's Reading Scale, Alpha 2, Adapted to Individual Testing"; in *Teachers College Record*, vol. xviii, pp. 253-60 (May, 1917).

² Kelley, T. L. "A Simplified Method of Using Scaled Data for Purposes of Testing"; in *School and Society*, vol. iv, pp. 34-37.

on which the pupil may be expected to work with an accuracy of 50 per cent. This procedure will then be recognized as simply calculating the median of the distribution of the

TABLE II. KELLEY'S SHORTER METHOD OF CALCULATING DIFFICULTY SCORES APPLIED TO TRABUE'S LANGUAGE-COMPLETION TEST

<i>I</i> Level of difficulty	<i>II</i> Number right	<i>III</i> Decrease in number right	<i>IV</i> Midpoints of levels of difficulty	<i>V</i> Products of <i>III</i> and <i>IV</i>
2.00	10			
4.00	10	..	3.0	..
6.00	10	..	5.0	..
7.00	10	..	6.5	..
8.00	10	..	7.5	..
9.00	10	..	8.5	..
10.00	8	2	9.5	19.0
11.00	9	-1	10.5	-10.5
12.00	7 Est	2	11.5	23.0
13.00	5 Est	2	12.5	25.0
14.00	3 Est	2	13.5	27.0
15.00	1 Est	2	14.5	29.0
16.00	0 Est	1	15.5	..
	..	10)128.0(12.8

differences. If some standard of accuracy other than 50 has been chosen, the procedure must be changed accordingly.

The above method has been modified by Kelley by using the average instead of the median of the differences. This has been done in Kelley's arrangement of Trabue's Completion-Test Exercises for individual testing. His procedure is illustrated in Table II. In the first column the levels of difficulty of the scale are given. In the second column the record of a pupil is given in terms of the number of exercises done correctly on each level. Beginning with level 12 the scores are estimates made in accordance with the relation expressed by the ogive curve. In column three the decrease in number right is given for successive levels of difficulty. In column four the midpoints of the successive levels of difficulty are given. The last column contains the products of columns three and four. To obtain the pupil's score, the algebraic sum of these products is divided by 10, the number of units on each level of difficulty. The quotient is 12.8, which defines the level of difficulty on which this pupil would be most likely to give a performance which was exactly 50 per cent accurate. It should be noted that the decrease in the accuracy of the pupil's performance from level 10 to level 11 is -1 . This decrease is simply considered as a negative quantity and does not in any way complicate the procedure.

Van Wagenen's shorter method of estimating scores. Van Wagenen¹ has further simplified the procedure for calculating a pupil's score on a scaled performance test by providing tables from which the estimates may be read directly. His tables have been constructed for 10 exercises or tasks on each level of difficulty, and for levels representing integral difficulty values. Similar tables could be constructed for other scales. When a pupil's performance includes a level

¹ Van Wagenen, M. J. "Table for Computing Mean Individual Scores in Educational Scales"; in *Teachers College Record*, vol. xxi, pp. 441-51 (November, 1920). See also his reading scales in history and general science.

of difficulty on which he has made no errors, the procedure for calculating his score is simple. Take the lowest level of difficulty for which the number of errors is 10. From Van Wagenen's table one can obtain the estimated level of difficulty on which the number of errors would be 5 or 50 per cent of the number of possible errors. If there is no level for which the number of errors is 10 — i.e., if the pupil has done some of the exercises correctly on the highest level of the scale — the table provides for making a similar estimate from his record on this level. This estimate is not the pupil's score. It must be corrected for all errors made on lower levels of the scale. The method of doing this is to find the sum of the errors made on these levels. Divide this sum by 10, and subtract the quotient from the estimate.¹ The remainder is the pupil's score; i.e., the theoretical level of difficulty on which he would just be able to do 50 per cent of the exercises correctly. If the pupil fails to do correctly all of the exercises on the lowest level of the test, it is necessary to estimate the number of additional errors he would have made if the test had been extended downward to such a level. These estimates are given in a table and must be subtracted from the corrected estimate obtained above.

Description of scaled performances in terms of the number of exercises done correctly. The plan of describing a pupil's total performance in terms of the level of difficulty on which he has just barely attained a given standard of accuracy requires that the performance on each level of difficulty be described with reference to accuracy.² To do this with even moderate precision requires that a pupil do

¹ This division by 10 may be avoided by multiplying the scale values by 10, and thus making the unit .1 *P.E.* This Van Wagenen has done in his reading scales for history and general science.

² When there are the same number of exercises on each level this description may be given in terms of the number of correct answers.

several exercises on each level of difficulty or that a system of partial credits be devised. The inclusion of a large number of exercises in a test will obviously require a large expenditure of time in testing. In order to reduce the testing time test-makers in certain fields have included only one exercise on each level of difficulty. This has been done by Woody, Starch, Hotz, and others. Even when the pupil is asked to do several exercises on each level of difficulty, as in the Thorndike Scale, Alpha 2, some users of such tests have described the pupil's performance in terms of the number of exercises which he does correctly.

It may be contended that this procedure amounts to giving equal credit for exercises which explicitly vary widely in difficulty. However, if the exercises are equally spaced along the scale of difficulty, and the easiest one has a scale value of 1, the number of exercises done correctly will give an approximate description of the level of difficulty reached. Even when these two conditions are not met, the description may be sufficiently accurate for practical purposes.

Relation between number of exercises done correctly and score in terms of level of difficulty. Van Wagenen's General History Scale A and General History Scale B were given to a group of pupils. The test papers were scored both according to the directions given by Van Wagenen and also in terms of the number of exercises answered correctly.¹ Scale A was given to 36 pupils and Scale B to 34. The coefficient of correlation between the regular scores and the number of exercises done correctly was found to be $.87 \pm .028$ for

¹ In this test the pupil is asked to check a list of statements giving those that are in agreement with a paragraph that he has read. A pupil may make errors in two ways. He may check a statement that is not true, and he may fail to check a statement that is true. He may also do an exercise correctly in two ways, either by checking true statements or by not checking statements which are untrue.

Scale A, and $.91 \pm .021$ for Scale B.¹ For Scale A the probable error of estimate in terms of the regular score is 2.9.² In terms of the number of exercises done correctly it is 2.1. For Scale B the probable errors of estimate are 2.6 and 2.1, respectively. Thirty-one of these pupils took both Scale A and Scale B. The coefficient of correlation between the two sets of the Van Wagenen scores is $.78 \pm .04$. The probable variable error of measurement³ is 1.2. Although the probable error of estimate is approximately twice as large as the probable error of measurement, it is not likely that the description of a pupil's performance on this test, in terms of the number of exercises done correctly, introduces serious errors. Considering the magnitude of the probable variable error of measurement, it would appear that one would be entirely justified in using the simpler method of describing the pupil's performance.

The sum of difficulty values used as scores. It has been proposed that the sum of the difficulty values of the exercises done correctly should be taken as a pupil's score on a scaled performance test. This procedure results in a pupil receiving more credit for doing a difficult exercise than for doing an easy one. In the Virginia Survey a spelling test was constructed by taking one word from each of the twenty consecutive columns of the Ayres Spelling Scale. Thus words which gradually increased in difficulty were secured. These words were weighted by assigning a value of one to the first, of two to the second, and so on, with the twentieth word having a value of twenty. The pupil's score was taken as the sum of the weights of the words spelled

¹ For an explanation of the statistical terms used here and in the following pages, see Chapters IX, XII, and XIII.

² The formula for the probable error of estimate is

$$P.E._{Est.} = .6745 \sigma_s \sqrt{1 - r_{12}^2}$$

³ This is computed by the formula, $P.E._M = .6745 \sigma \sqrt{1 - r_{12}^2}$

correctly. Under the writer's direction, this spelling test was given to 306 fifth-grade pupils in one school system. The performances of the pupils were described according to the rules used in the Virginia Survey. They were also described by giving one credit for each word spelled correctly. This assumes that the same credit should be given for the different words. The coefficient of correlation between weighted scores and unweighted scores for this group was $.96 \pm .0038$. In the same school system the test was given to 184 seventh-grade pupils, and the coefficient of correlation between the weighted and unweighted scores for this group was found to be $.97 \pm .0050$.

The coefficient of correlation expresses only a general relationship. These coefficients approach perfect correlation closely, and indicate that there is close agreement between the two sets of scores. We may also express the closeness of this relation in terms of the probable error of estimate. For the fifth-grade group it is 10.4 in terms of the weighted scores and .57 for the unweighted scores. The standard error of estimate for the unweighted scores means that, if the weighted scores were reduced to the same scale as that of the unweighted scores, so that equivalent scores would be identical in magnitude, fifty per cent of the pupils would receive scores which differ from each other by no more than .57 of a word. In the seventh grade the standard errors of estimate are 10.2 and .50. Thus, it is seen that in this case it makes very little difference which of these two methods of scoring is used. The error introduced by giving the same credit for all words is slight. Incidentally, it may be noted that the reliability coefficient of this test is probably no greater than .75 for these groups of pupils. For this degree of reliability, which is high for a twenty-word test, the probable error of measurement is approximately 10.0 for the weighted scores, and 0.5 for the unweighted.

Thus, the difference between the two methods of scoring is no greater than the probable error of measurement.

Combined scores. For certain purposes it is desirable to have the pupil performance described in terms of a single number. This necessitates that we select only one dimension of his performance, or that we combine the descriptions of the dimensions which are recognized. The number of exercises done correctly has been used as the pupil's score on rate tests. On such tests this score is a combination of rate of work and accuracy of performance. For example, a pupil may make a score of ten exercises right by doing ten exercises with 100 per cent accuracy. He may also make a score of ten exercises right by doing twenty exercises with 50 per cent accuracy. The number of exercises done correctly is essentially the product of the rate of work and of the accuracy when the latter is expressed as a per cent. When the number of exercises done correctly is used to describe a pupil's performance on a scaled test, his score is a combination of quality and difficulty. If the scaled test is timed, the rate of work will be included in the combination. In his Oral Reading Test, W. S. Gray uses a plan for securing the combination of three dimensions. A pupil is given a single score which is a function of his rate of reading, the quality of his reading, and the difficulty of the exercises which he has read.

A combined score is useful when a summary or general description of the pupil's performance is desired. However, the number of exercises right is an unanalyzed score and for this reason its interpretation is limited. On a rate test we cannot know whether the pupil's deficiency is in rate or in quality. When difficulty is introduced the situation is even more complex.¹

Accuracy when number of answers is limited. When the

¹ See page 87 for the "law of the single variable."

number of possible answers is limited there is a chance that a pupil will give the correct answer by chance. This is especially true in the case of exercises which permit of only two answers. In general a pupil who knows nothing about the exercises will be able to answer half of them correctly by guessing. It, therefore, becomes necessary to make allowance for this condition in describing the quality of a pupil's performance on such a test. When there are as many as four or more possible answers the effect of this condition is slight and is usually neglected in computing the pupil's score. The following formula gives a method for making allowance for a limited number of possible answers. The number of possible answers is represented by n , the number of answers correct by c , and the number of answers wrong by w . A pupil's score equals

$$\frac{c - w}{n - 1}$$

When there are only two possible answers, as in a "true-false" test, this formula becomes "the number right minus the number wrong."

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What information must be given in order to describe a pupil performance completely?
2. What are the requirements for describing the rate of a pupil's performance?
3. What is the usual scale of accuracy?
4. How is accuracy defined? What are the requirements for describing a performance with reference to accuracy?
5. What determines the standards of accuracy?
6. In a scaled performance a typical pupil does not do all exercises correctly up to a certain point and then fail to do any exercises beyond that point. A typical performance is one in which the pupil fails upon some exercises and then does correctly other exercises which are more difficult. Explain why this happens.
7. Upon what basis can the use of the number of exercises done correctly as a pupil's score on a difficulty scale be justified?

8. What is the application of the "law of the single variable" with reference to the description of a pupil performance?
9. In terms of what dimensions is the pupil's performance described in the following tests:
 1. Gray's Oral-Reading Test?
 2. Woody's Arithmetic Exercises?
 3. The Cleveland Survey Tests in Arithmetic?
 4. The Courtis Standard Research Tests in Arithmetic, Series B?
 5. The Van Wagenen Reading Scales in History and General Science?
 6. Courtis's Silent-Reading Test, No. 2?
 7. Minnick's Geometry Test?

SELECTED REFERENCES

BURGESS, MAY AYRES. *The Measurement of Silent Reading*, pp. 32-47. Russell Sage Foundation Monograph, 1921.

GRAY, WILLIAM SCOTT. *Studies of Elementary School Reading through Standardized Tests*. Supplementary Educational Monographs, University of Chicago, vol. I, No. 1, 1919.

KELLEY, T. L. "A Simplified Method of Using Scaled Data for Purposes of Testing"; in *School and Society*, vol. IV, pp. 34-37 (1916).

KELLEY, T. L. "Individual Testing with Completion-Test Exercises"; in *Teachers College Record*, vol. XVIII, pp. 371-82 (September, 1917).

KELLEY, T. L. "Thorndike's Reading Scale, Alpha 2, Adapted to Individual Testing"; in *Teachers College Record*, vol. XVIII, pp. 253-60 (May, 1917).

MINNICK, J. H. "A Scale for Measuring Pupils' Ability to Demonstrate Geometrical Theorems"; in *School Review*, vol. XXVII, pp. 101-09 (1919).

MINNICK, J. H. *An Investigation of Certain Abilities Fundamental to the Study of Geometry*. University of Pennsylvania Monograph, 1918.

STARCH, D. "A Scale for Measuring Ability in Arithmetic"; in *Journal of Educational Psychology*, vol. VII, pp. 213-22 (April, 1916).

VAN WAGENEN, M. J. "Table for Computing Mean Individual Scores in Educational Scales"; in *Teachers College Record*, vol. XXI, pp. 441-51 (November, 1920).

WOODY, CLIFFORD. *Measurements of Some Achievements in Arithmetic*. Teachers College Contributions to Education, No. 80.

CHAPTER VI

DESCRIPTION OF PERFORMANCES OF PUPILS: QUALITY SCALES

Description of a performance when quality does not mean accuracy. When a pupil's performance or some division of it cannot be described as right or wrong, or at least as partially right, it is necessary to construct a scale which can be used in describing its quality. This condition exists in handwriting, English composition, hand-sewing, and drawing. A quality scale is simply a collection of sample performances of the type to be described, which exhibit varying degrees of quality from the lowest quality with which we have to deal up to the highest such quality. The samples are arranged in order of increasing quality, and usually the quality of each is described numerically. It is desirable to have the samples selected so that the differences in quality between successive samples are approximately equal. It is also desirable to have the numerical designations of the degrees of quality express the amount of quality measured from a zero which means no quality at all.

A quality scale of this type is used by matching the performance to be described with the sample of the scale which most nearly resembles it in quality. The decision which the scorer reaches is only an opinion, and we find that different persons match the same performance with different scale samples. Hence this matching is a subjective process; i.e., it depends upon the person who does it. Appropriate training will materially reduce the subjectivity of the process.¹

¹ Thorndike, E. L. "Teachers' Estimates of Specimens of Handwriting"; in *Teachers College Record* (November, 1914).

Gray, C. T. "The Training of Judgment in the Use of the Ayres Scale of Handwriting," in *Journal of Educational Psychology*, vol. vi, pp. 85-97 (February, 1915).

Construction of a quality scale. The essential steps in the construction of a quality scale are three:

1. Collecting sample performances which represent as wide a range of the quality to be described as is possible.

2. Selecting from this collection a limited number of samples (10 to 20) which represent the entire range of quality, and which differ from each other by approximately equal increments of quality.

3. Determining the quantitative description of the quality of each sample with reference to an established zero point, and arranging the samples in the form of a scale.

1. Collecting sample performances. In collecting sample performances for the purpose of constructing a quality scale, it is necessary that all of the performances be of the same type. For example, if samples of hand-sewing are being collected, all pupils should attempt to do the same kind of stitches with the same material and equipment. If samples of handwriting are being collected, all pupils should write the same sentences under the same conditions. There should be some performances approximating zero in quality, and there should be others which exhibit the highest possible degrees of excellence. All degrees of quality between these two extremes should be represented. In collecting samples of certain types of hand-sewing from which to construct a scale for the measurement of the quality of hand-sewing, Dr. Murdoch secured samples not only from school children, but from some adults who were generally recognized as being expert needlewomen, and also from the inmates of an institution for the feeble-minded. This procedure resulted in a wide range of quality. The number of samples collected is unimportant, except that this number should be sufficiently large so that it is reasonably certain that all degrees of quality will be represented. Most recent makers of scales have started with at least a thousand samples in their original collections.

2. *Selecting samples to form a scale.* In selecting from all of those collected a set of samples to form a scale, the objective to be realized is a small number of samples (10 to 20) which will be representative of the entire range of quality from the poorest sample collected up to the best sample collected. It is desirable that the differences in quality between successive samples be approximately equal. Any plan of selection which will accomplish this purpose is justified. The best procedure is to make a series of selections. For example, one may start with 2500 samples of handwriting. These samples should be numbered for purposes of identification. The first selection may be made by a single judge who sorts the 2500 samples into 10 piles, representing 10 different degrees of quality of handwriting. Any sample which is incomplete or otherwise unsatisfactory for scale purposes should be discarded. After the first sorting is completed the judge should examine each pile, and make any changes which seem desirable. Then he should select 25 or 30 samples from each pile which will be representative of the pile. This will reduce the number of samples to nearly 300, but they should be representative of the original collection.

From this derived collection a selection of 50 or 60 samples should now be made. In doing this the independent judgment of a number of competent judges should be secured. Each judge should be asked to sort the 300 samples into 10 piles. They should be instructed to make the differences in quality between these successive piles approximately equal. The piles should then be numbered 1, 2, 3, 4, . . . 9, 10, beginning with the pile representing the lowest degree of quality. A record sheet similar to that illustrated in Table III will be needed for recording the results of the sorting by each judge. After each sorting the samples should be thoroughly shuffled, so that no judge will be influenced by the

sorting of any other. The totals represent the consensus of opinion of these judges with reference to the quality of the respective samples. The sample which has the smallest total is the one which is judged to possess the least quality.

TABLE III. RECORD SHEET FOR RECORDING JUDGMENTS OF JUDGES CONCERNING THE QUALITY OF SAMPLES

<i>Number of sample</i>	<i>Judge</i>						<i>Total</i>
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>		
1	8	7	5	9	7		36
2	3	2	1	1	2		9
3	4	6	4	5	4		23
4	7	5	8	7	6		33
5	10	9	10	8	10		47
6							
7							
8							
9							
10							

The sample which has the largest total is judged to represent the highest degree of quality. It is wise to secure the judgments of 10 to 30 judges at this stage of the selection.

By referring to the totals, 50 to 60 samples should now be selected, which are representative of the entire range of quality. These samples should then be submitted to a still larger number of judges and the procedure described above repeated.

Using the totals for the 50 to 60 samples as a basis, it should now be possible to select the samples intended for use in the final scale. Usually, the selection will be made by taking the sample having the smallest total and the sam-

ple having the largest total, and those samples which appear to represent equal increments of quality between these two extremes. The number chosen at this time should be not less than 10, nor more than 20. Occasionally, it may be desirable to have more than 20 samples. It is possible that a few samples chosen at this time may later prove unsatisfactory and it will be necessary to reject them. It is, therefore, wise to select a few more samples than are to be included in the final scale.

The process of selection by successive approximations can be greatly extended when desired. By increasing the number of judges used at each stage, and by increasing the number of successive selections, the process of making a quality scale can be greatly elaborated. Scale-makers appear to have secured satisfactory results by making three or four successive selections. It is, however, necessary that competent judges be used at each stage of the work. In our illustration, the first selection was made by one judge. Some test-makers have used from five to ten judges for this first step. This greatly increases the labor and does not appear to be necessary. For the final selection the judgment of forty to fifty competent persons should be sought.

The test of the authenticity of the selection is in the determination of the numerical values which represent the degrees of quality of the samples chosen. If these numerical values are such that the samples differ from each other by approximately equal amounts of quality a good selection has been made. If the differences are not approximately equal the selection was not a good one.

3. *Determining the quantitative description of the quality of the samples.* Three methods have been used for determining the quantitative descriptions of the samples which have been chosen for a quality scale. Two of these methods, which may be called subjective, were described by Thorn-

dike in connection with the derivation of his handwriting scale, which was the pioneer measuring instrument of this type. The first method is to ask competent judges to assign numerical values to the samples, such as 1, 2, 2.7, 3, 3.4, 4.0, etc. The average of the values assigned to a sample by a large number of competent judges is then taken as the numerical description of the quality of the sample. When a value of 1 is assigned to the sample it is judged to possess the least quality. It is assumed that this quality is just one unit above zero. This assumption may or may not be justified in a particular case. When it is not, the values determined by this method will need to be shifted upward or downward, depending upon the location of the zero that is finally agreed upon. The method possesses another limitation in that competent judges are unable to discriminate between very poor samples to the same extent that they are able to discriminate between samples possessing moderate quality. The same is true of very good samples. This limitation, of course, could be removed if we were able to include in our final selection a number of samples at each end of the scale which were later discarded.

Modifications of the method. This method may be modified when revising an existing scale or when constructing a scale in a field where there is already a scale.¹ Instead of securing estimated values, the samples may be rated on the existing scale by a large number of competent judges. The average or median of their ratings would be taken as the scale values of the samples.

The second method requires that the samples be ranked in order of quality by each judge. This ranking should be described with reference to some arbitrary ranking which will approximate the true ranking. Probably the best rank-

¹ Trabue, M. R. "Supplementing the Hillegas Scale"; in *Teachers College Record*, vol. xviii, pp. 51-84 (January, 1917).

ing for this purpose is the one according to the totals on the basis of which the final selection of samples was made. In Table IV, which illustrates the method of recording the rankings, the samples beginning with the highest were ranked

TABLE IV. FORM OF RECORDING THE RANKINGS OF SAMPLES BY JUDGES IN ORDER TO SECURE PER CENT OF "BETTER JUDGMENTS"

No. of sample	Judges										Total above	Per cent above
	I	II	III	IV	V	VI	VII	VIII	IX	X		
79	A	B	A	A	A	B	B	A	A	A	7	70
84	B	A	A	B	A	A	A	B	B	A	6	60
42	A	A	A	B	A	A	A	B	A	A	8	80
108	B	A	A	A	B	A	A	A	A	B	7	70

as shown in the first column. Sample number 79 had the highest total in the final selection. Sample number 84 came next, and so on. When arranging the samples in order of merit judge I ranked sample number 79 above sample number 84. This is shown by the letter A. This judge ranked sample number 84 below sample number 42. This fact is shown by a B. After the rankings of all judges have been recorded, the total number of judges who ranked each sample above the one just below it in the arbitrary ranking should be entered in the next to the last column. In the last column the per cent of "better judgments" is given. This is obtained by dividing the numbers in the next to the last column by the total number of judges. From the per cent of better rankings the numerical description of the quality of each sample can be calculated.

Table V has been constructed from data given by Dr. Murdoch in her description of the construction of her scale for measuring certain elements of hand-sewing. The first

TABLE V. DIFFERENCES IN QUALITY CORRESPONDING TO PER CENT OF "BETTER JUDGMENTS"

(Arranged from Tables 2 and 5 of Dr. Murdoch's monograph)

<i>Identification number of sample</i>	<i>Per cent of judges placing sample above next sample</i>	<i>Difference in quality in terms of P.E.</i>
544	42	-.299
977	88	1.742
652	76	1.046
976	56	.224
900	64	.532
381	68	.694
322	68	.694
653	66	.612
709	66	.612
321	78	1.143
497	58	.299
671	80	1.246
511	60	.376
473	72	.865
678	56	.224
418	56	.224
448	82	1.355
440	60	.376
489	80	1.246
528	84	1.472
532	84	1.472
530		

column gives the identification numbers of the 22 samples in the final selection. The second column gives the per cent of 50 judges placing each sample above the next sample. In the third column the differences in quality between successive samples are given in terms of *P.E.* These differences are obtained from a table similar to Table I.¹ If we look in this table for 42 per cent, we find that the *P.E.* equivalent for 42.07 per cent is .30 *P.E.* In Table V the difference in quality between sample number 544 and 977

¹ See page 96.

is given as $-.299$. The procedure followed in ranking the samples is such that the signs of the *P.E.* values are just the opposite of those given in Table I. With this change the differences in quality in terms of *P.E.* may be read directly from Table I.

This method of determining the quantitative description of the quality of samples is based upon the theorem that differences in quality which are noticed equally often by competent judges are equal in magnitude unless they are noticed by all such observers, or are noticed by only 50 per cent of the observers. The method also assumes that the errors in the judgments expressed by an individual judge form a normal distribution about zero, and that the variability of this distribution is the same for all individuals and all samples.

Location of zero point. The location of the zero point is usually arbitrary. Dr. Murdoch in constructing her scale for measuring certain elements of hand-sewing was fortunate in securing a sample of hand-sewing for which she obtained an average value of approximate zero when the opinions of a number of competent judges were secured in regard to its quality. When a scale-maker is not fortunate in having a sample which the judges agree upon as representing zero, it is necessary to secure the consensus of their opinions concerning the distance of certain samples above zero.

The subjectivity of these methods. Both of these methods for determining numerically the descriptions of the qualities of the samples chosen for a scale are subjective, in that they are based upon the opinions of judges. It should, however, be noted that competent judges are to be used in every case, and that the values represent the consensus of opinion or average estimate of these judges. The procedure may be justified by pointing out that this is exactly the way quality is defined. The quality of a painting is exactly what com-

petent judges consider it to be. As it is usually thought of, the quality of handwriting is defined in terms of the consensus of opinion of competent persons. Hence, the fact that the methods just described are based upon opinions does not constitute a valid criticism.

Objective methods of scale construction. Some scale-makers, notably Ayres and Freeman, have attempted to secure some objective index of quality. In constructing his handwriting scale, Ayres defined legibility, which was the quality he considered, in terms of the average rate at which the samples could be read. He assumed that, the other factors being constant, the legibility of a sample of handwriting was inversely proportional to the number of seconds required for reading it. He therefore had a large number of samples, which he collected, read by a number of trained persons under carefully controlled conditions. From the data thus obtained he calculated a numerical index of the legibility of each sample. He then selected a group of samples which differed from each other by approximately equal amounts of legibility, and which for other reasons seemed suitable for being included in a scale. He assumed that the numerical indices of legibility for an unselected group of samples would form a normal distribution. The base line of this normal distribution from -2.5σ to $+2.5\sigma$ was divided into one hundred parts.¹ Samples were then picked out to represent the numerical values of 20, 30, 40, etc.

Freeman took such qualities as uniformity of slant, uniformity of alignment, and so forth, and attempted to secure objective descriptions of these characteristics of handwriting. He used these objective descriptions in much the same way that Ayres used his data in constructing his scale.

¹ Ayres's account of this scale does not state definitely that the extent of the base line was taken as 5σ . However, there is evidence which points to this choice.

Using a quality scale. The description of performances by means of a quality scale is subjective; i.e., it depends upon the person giving it. Hence, it is necessary to exercise care in order to reduce the error of measurement due to subjectivity to a minimum. This is accomplished by following a systematic procedure by training. Most scales are accompanied by detailed directions for their use. These should be followed. Two or three hours of practice will also reduce the degree of subjectivity.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What is a scale of quality?
2. Quality scales are needed in the description of what types of performances?
3. What are the essential steps in the construction of a quality scale?
4. What criteria should govern one in a collection of samples for the construction of a quality scale?
5. What are the two subjective methods for selecting and evaluating samples for a quality scale?
6. What assumptions are involved in each method?
7. What are the sources of difficulty in securing accurate measurements of quality by means of quality scales?
8. How did Ayres construct his scale of quality in handwriting? Is his method preferable to the one used by Thorndike? Give reasons.
9. Which is the best general handwriting scale? Look up the experimental evidence.

SELECTED REFERENCES

AYRES, L. P. *A Scale for Measuring the Quality of Handwriting of School Children*. Bulletin No. 113, Division of Education, Russell Sage Foundation, New York City.

GRAY, C. T. "The Training of Judgment in the Use of the Ayres Scale for Handwriting": in *Journal of Educational Psychology*, vol. vi, pp. 85-97 (February, 1915).

MURDOCH, KATHARINE. *The Measurement of Certain Elements in Hand-Sewing*, chap. III. Teachers College Contributions to Education, No. 103.

STARCH, DANIEL. "A Scale for Measuring Handwriting"; in *School and Society*, vol. ix, pp. 154-58 and 184-88 (February 1 and 8, 1919).

STARCH, DANIEL. "A Revision of the Starch Writing Scale"; in *School and Society*, vol. x, p. 498 (October 25, 1919).

STARCH, DANIEL. "Methods in Constructing Handwriting Scales"; in *School and Society*, vol. x, p. 328 (September 13, 1919).

THORNDIKE, E. L. "Handwriting"; in *Teachers College Record*, vol. II No. 2 (March, 1910).

TRABUE, M. R. "Supplementing the Hillegas Scale"; in *Teachers College Record*, vol. xviii, pp. 51-84 (January, 1917).

CHAPTER VII

DERIVED SCORES

The lack of a common unit of measurement. A variety of units are used in describing the performances of pupils according to the methods considered in the two preceding chapters. The unit of work, in terms of which the rate of a pupil's performance is described, is generally determined by the structure of the test. Consequently, different tests are likely to define different units of work. Even in the case of tests where the units are called by the same name, they are likely not to be equal in magnitude. For example, in the Courtis Standard Research Tests, Series B, an example is taken as a unit of work; but an example in the addition test is not equal to an example in the subtraction test. Each of the four tests in this series has a different unit, although they are called by the same name. Accuracy is generally measured on a scale which ranges from zero accuracy to perfect accuracy. The unit is one hundredth of this range. However, both zero accuracy and perfect accuracy depend upon the standard of accuracy which is adopted for a given test. Since standards of accuracy vary for different tests, it follows that the unit of accuracy also varies. Hence, in the description of both rate and accuracy, different units of measurement are likely to be employed with different tests.

The unit used in describing a level of difficulty which a pupil or class reaches on a scaled test is a measure of the variability of the ability of a certain group of pupils. The median deviation of the variability (*P.E.*) is one measure which has been used as a unit by a number of test-makers, such as Woody, Trabue, Hotz, and Henmon. Other work-

ers have used the standard deviation (σ) or some fraction of it as the unit. The constancy of a measure of variability depends upon the definition of the group of pupils whose abilities are considered. Grade groups have been most frequently used, but a given grade group, such as the fifth grade, is not well defined. The practice with reference to the classification and promotion of pupils is not uniform. In some school systems a distinct effort is made to classify the pupils so that the variability with respect to the abilities engendered within the class is reduced to a minimum. In other school systems pupils are classified on a different basis and, as a result, the variability of those belonging to a given school grade in such a school system is much greater. Thus two test-makers utilizing different school systems in the construction of similar tests might define units of difficulty differing in magnitude. Furthermore, the variability appears to increase slightly as we advance from a lower to a higher grade. If this is the case, the size of the unit would depend upon the school grade that is chosen to define it. The median deviation (*P.E.*) of the third grade would not be equal to the *P.E.* of the eighth grade. Hence, a measure of the variability of grade groups does not furnish us with a constant unit for describing the performances of pupils.

The lack of a common zero point. Not only are different units used in the description of the performances of pupils, but the descriptions of quality and level of difficulty reached are likely to be expressed with reference to different zero points. The standard of accuracy adopted establishes the zero point, and, as we have already shown, the standards used are arbitrary. Hence, the measures are expressed from arbitrary zero points and do not necessarily mean no accuracy. The zero points of quality scales are more uniform in meaning. In the case of scaled tests, the makers have generally attempted to establish zero points which approximate ab-

solite zero, and which might be interpreted as meaning the absence of ability or not any of the ability being measured. It is, however, not certain that they have succeeded in their attempts, and therefore it is likely that there is little uniformity in the zero points which have been established for scaled tests.

In the case of the rate of work, zero means not any rate, or that the pupil failed to do a single unit of work within the time allowed. However, zero rate cannot be interpreted as meaning the possession of zero ability. A pupil may possess some ability in a given field, such as silent reading, spelling, or the operations of arithmetic, and yet be unable to do a single unit of work within the time allowed because the exercises presented to him are more difficult than he can do. For this reason, if measures of rate of work are thought of in terms of ability rather than merely as the pupils' rates of work, we have these measures also expressed with reference to different zero points.

The need for a common unit and a common zero point. Since no two tests employ the same unit or the same zero point, except by chance, the scores used to describe the performances of pupils are not comparable. A score of 19 on one test does not, except by chance, mean the same thing as the score of 19 on another test. Considerable inconvenience results from this condition. Such point scores are without meaning until they are compared with norms. Since different units and different zero points are used, this makes necessary a different set of norms for each test, even in the case of tests designed to measure the same abilities. This adds to the labor of test construction, and makes it necessary for the test user to have at hand a statement of the norms for all tests with which he is dealing. It is not possible to combine the scores obtained from two or more tests, or to compare them in any precise way, without employing an

elaborate statistical procedure to reduce them to the same scale. The recording of scores on permanent record cards is also complicated. It is necessary to make a record not only of the scores but also of the norms. Otherwise, without first looking up the norms, there can be no interpretation or comparison of the scores which pupils make at successive periods of testing.

Derived scores in terms of a common unit and a common zero point. The description of a pupil's performance according to the rules which accompany a test may be called a point score. In order to secure a common unit and a common zero point a number of plans have been proposed for translating the point scores into derived scores. We shall consider certain of these proposals.

The variability of chronological-age groups as a common unit. As indicated above, a measure of the variability of grade groups does not furnish us with a constant unit, because the definition of such groups is not fixed. In order to avoid this limitation, McCall has proposed that we use the group of pupils whose chronological ages fall between twelve and thirteen years. Since practically all pupils of these ages are to be found in our public schools, twelve-year-old pupils would always be selected no matter from what school system they were taken. Such a group would always possess the same average general intelligence, provided it was a random selection from a typical population.

The character of the instruction which the pupils had received would doubtless affect their variability slightly. It has been pointed out that the usual plan of instruction does not furnish the bright pupils with adequate opportunities for learning. The emphasis has been placed upon the instruction of backward and dull pupils. If our procedure were modified so that backward and dull pupils received less attention, and the bright pupils received more, it might well

be that the variability of the chronological-age groups would be affected. Furthermore, we do not know that the variability is the same for different abilities. For example, the variability might be greater for the ability to add than for the ability to subtract. The variability of the ability to read might differ from that of ability in the field of music.

Another objection to the universal application of this proposed unit is that some abilities have not been acquired by all twelve-year-old pupils. This would be true of many abilities engendered in the high school, because most twelve-year-old children have not yet entered high school. In the case of tests designed for use in the primary grades it would be necessary to introduce some modifications when applying them to twelve-year-old pupils, most of whom would be found in the intermediate and grammar grades. In many cases the measuring instruments would not be suitable for use with most twelve-year-old pupils and, consequently, would fail to yield a measure of their abilities. It has been suggested that this limitation might be overcome by using sixteen-year-old pupils for abilities engendered in the high school, and eight-year-old pupils for measuring instruments designed for use in the primary grades. This suggestion appears to be feasible in the case of some tests designed for use in the primary grades, but for abilities engendered in the high school a difficulty is introduced by reason of the fact that there is a differentiation of the students with respect to courses studied. All pupils do not study the same school subjects, and there is evidence to show that we do not secure a random sampling of pupils in the different school subjects. For example, one investigation indicates that pupils studying Latin possess a higher average general intelligence than do pupils studying commercial subjects.

However, in spite of these limitations, the proposal to use the variability of the ability of twelve-year-old pupils will

result in a unit which is approximately the same for a number of tests. Whenever the conditions are such that it can be used, a grouping of pupils on this basis is to be preferred to a grouping on the basis of school grade.

A proposed common zero point. A common zero point is also necessary. It is generally difficult or impossible to determine an absolute zero point which will mean not any of the ability in question. Therefore, in order to secure uniformity, McCall has proposed the use of a zero point 5σ below the average ability of twelve-year-old pupils. He shows that this zero point approximates the determinations of absolute zero points made for a number of particular tests. Even if such a zero point does not represent absolute zero of ability, it will be useful because it will always have a meaning of being just 5σ below the average ability of twelve-year-old pupils.

McCall's derived scores. McCall¹ has used a unit $.1\sigma$ of the distribution of the abilities of all twelve-year-old pupils. This unit he calls "T," after Thorndike and Terman. His scale extends from 5σ below the average of the ability of twelve-year-old pupils up to a point 5σ above this average. Thus, he has a scale of 100 T units. A pupil's performance is first described in terms of a point score. In the case of the Thorndike-McCall Reading Scale, in connection with which McCall described his plan for derived scores, this point score is the number of questions answered correctly. The T score which is equivalent to each point score is determined by ascertaining, for the group of twelve-year-old pupils, the per cent of pupils exceeding each point score.² The

¹ McCall, W. A. "A Proposed Uniform Method of Scale Construction"; in *Teachers College Record*, vol. xxii, pp. 31-51 (January, 1921).

² More exactly, this is the per cent exceeding plus half of those reaching each point score. The number of pupils making a given point score is divided, half being grouped with those making higher scores and half with those making lower scores.

relation between the per cent of pupils exceeding given point scores and the corresponding T scores is given by Table I.¹ For example, McCall found, in the case of the Thorndike-McCall Reading Scale, that 93.1 per cent of the twelve-year-old pupils exceeded a point score of 12. By referring to Table I, we find that 93.1 per cent of the pupils in a normal distribution, which is assumed for the twelve-year-old group, are above a point 1.50σ below the average of this group. If the position of this point is expressed with reference to a zero point taken 5σ below the average it would be $+3.5\sigma$. Since T , the unit used by McCall, is $.1\sigma$, the position of this point would be described by the score $35T$. Therefore, a point score of 12 on this test is to be translated into a derived score of $35T$ on McCall's uniform scale.

This plan of calculating derived scores may be applied to all tests which are appropriate for twelve-year-old pupils. It is only necessary to ascertain for each point score which the test yields the per cent of twelve-year-old pupils who exceed it. By means of Table I we will be able to ascertain the T score which is equivalent to any given point score. The adoption of this plan would mean that all scores would be expressed in terms of a common unit,² and with reference to a common zero point. This would result in the scores yielded by different tests having a similar meaning. For example, a score of 35 would always represent that degree of ability which was exceeded by 93.1 per cent of all twelve-year-old pupils. Similarly, a score of 62 on this scale would represent that degree of ability which is exceeded by only 10.6 per cent of all twelve-year-old pupils.

The advantage of using such derived scores is obvious. Since all scores are expressed in terms of the same unit, and

¹ See page 96.

² In the case of tests in different fields this assumes that the variability of the distributions of ability is the same.

with reference to the same zero, they may be added or subtracted. Only simple arithmetic is required to find a pupil's average score. A score of a given magnitude will always have the same meaning with reference to the norm for twelve-year-old pupils. The norm (average) for this group will always be 50. However, the necessity for separate sets of norms for different abilities, in the case of groups other than twelve-year-old pupils, will not be eliminated. Thus we should expect to have different sets of norms for different tests and to some extent for different plans of school organization and for different courses of study.

Pintner's derived scores. Pintner¹ has proposed a modification of McCall's procedure. Instead of expressing all scores in terms of the variability of one age group, he has prepared a separate scale for each age group. This plan has the advantage of assigning to each pupil a score which directly compares his ability with that of his own age group, but it does not facilitate the comparison between the abilities of pupils of different ages. Different units and different zero points are used for pupils of different ages. A common unit and a common zero point will be secured only for pupils belonging to the same chronological age group. The plan is obviously more complex than the one proposed by McCall and has less to commend it.

Derived measures expressed in terms of grade norms. It has been proposed that the point scores yielded by a given test be expressed in terms of the grade norms for that test.² For example, one set of proposed grade norms for the Courtis Standard Research Tests, Series B, in terms of the

¹ Pintner, Rudolph, and Marshall, Helen. "A Combined Mental-Educational Survey"; in *Journal of Educational Psychology*, vol. xii, p. 32 (January, 1921).

² Buckingham, B. R. "Suggestions for Procedures Following a Testing Program: I, Reclassification"; in *Journal of Educational Research*, vol. ii, p. 787 (December, 1920).

number of examples attempted by pupils just entering the respective grades are given in the table below:

<i>Grade</i>	<i>Addition</i>	<i>Subtraction</i>	<i>Multiplication</i>	<i>Division</i>
VIII	11.2	12.2	10.9	10.1
VII	10.3	11.1	9.6	8.9
VI	9.2	9.8	8.7	7.1
V	8.0	8.2	6.7	5.3
IV	6.2	6.2

If a pupil attempts 9 examples in addition he is shown to have a degree of ability between the norm for the fifth grade and the norm for the sixth grade. The norms, as given in the above table, are in terms of the abilities of pupils who are just entering the respective grades. Therefore, this pupil may be said to have a degree of ability ten twelfths greater than that possessed by the average of the pupils just beginning the fifth grade. According to this plan, this pupil's ability in addition could be represented by a derived score of 5.8.

This proposal for translating point scores into derived measures possesses very obvious limitations. In the first place, grade norms are not constant. They depend upon the organization of the school and the effectiveness of the instruction. If a measure of the variability of grade groups is unsatisfactory as a unit, certainly a scale consisting of the grade norms will be even more unsatisfactory. However, the most serious limitation is that the plan does not provide any means for translating either low scores or high scores into derived measures. Such a scale cannot extend below the norm for the lowest grade to which the test is given. It cannot extend above the norm for the highest grade to which it is given. Therefore, we have a scale which is inadequate at both ends. Furthermore, the plan presents

difficulties of application. For these reasons, derived measures in terms of grade norms should not be used. They have nothing to commend them except in the grade placement of pupils. Even in this activity they possess no essential advantage over other derived measures.

Derived measures expressed in terms of grade percentile scores. A pupil's performance may be described in terms of the position which it occupies in the standard distribution for his grade. This is accomplished by dividing the distribution so that each division contains the same per cent of the total number of scores. One plan is to make four divisions, or quartiles. A pupil's performance may then be described as being in a given quartile, such as the second quartile, or the fourth quartile. Other plans require a division of the distribution into 10, 20, or 100 parts. When it is divided into 100 parts each part is called a percentile, and a pupil's ability might be described as that of the 80 percentile or the 57 percentile. The objections which have previously been stated with reference to the instability of grade distributions apply to this proposal. There are also additional reasons why this plan is unsatisfactory. While the derived scores are expressed in terms that can be easily understood, they do not possess the qualities which are possessed by derived scores expressed in terms of the variability of the chronological-age group. The unit is not constant. In calculating percentile scores the distribution is divided so that equal areas are obtained. For this reason the divisions do not mark equal distances on the base line of the distribution. The result of this is that the difference between the degrees of ability represented by a 40 percentile score and a 45 percentile score is not at all equal to the difference between the degrees of ability represented by a 90 percentile score and a 95 percentile score. The latter is much larger.

Measures expressed as achievement ages. In using the Binet Scale for Measuring the Intelligence of Children, a pupil's performance is first described in terms of his success or failure in doing certain exercises. This description is then translated into a derived measure, called mental age. A pupil is said to have a mental age of nine years when the point score which describes his performance is equivalent to that of the average point score of nine-year-old children. Likewise, a pupil is said to have a mental age of twelve years when the point score that describes his performance is equivalent to the average point score of all twelve-year-old children. This plan of translating point scores into corresponding age scores has been applied to other intelligence tests. The procedure for obtaining the basis for making this translation has been to give the test to an unselected group of pupils for each age, and in this way determine the average point score which is made by the pupils of each age group. These chronological age norms then become the basis for translating point scores into the corresponding mental-age scores. The age group whose norm is equivalent to a pupil's point score becomes his mental age.

This procedure has been applied to achievement tests. McCall, Pintner, and others have determined the norms for chronological-age groups and used these as a basis for translating point scores into age scores. Since a chronological-age norm represents the average achievement of pupils of that age, it is necessary to introduce some modifications of this plan in order to provide for translating very low scores and very high scores into corresponding achievement ages. Pintner fails to make this provision. For example, the average of the highest group in Pintner's table is 102. The maximum point score is 141. McCall has assumed a rectilinear relationship between point scores and age norms. On this basis he has extended his age scale upward and downward

to provide for age scores corresponding to all point scores which his test yields.

The writer ¹ has proposed, in the case of arithmetic and silent reading, that mental-age groups be used instead of chronological-age groups in determining the basis for translating point scores into achievement ages. The reason for making this proposal is that mental age is a better criterion of ability in these two fields than chronological age.² Pupils having the same mental age are said to possess the same capacity to learn. However, since the average mental age of an unselected chronological-age group is numerically the same as the average chronological age, this difference in procedure will not produce a numerically different result. The use of mental-age groups has the advantage of certain logical considerations in connection with the interpretation of the resulting derived measures.

Quotients as derived measures. A score, even a derived measure of the kind we have described above, tells only the magnitude of a pupil's achievement. This has meaning only when compared with appropriate norms. In most uses of educational measurements we are concerned more with the comparison of the measure of a pupil's ability with an appropriate norm than with the absolute amount of his ability. For this reason, another type of derived measures have been proposed which would be an index of the relation of the measure of a pupil's ability to the appropriate norm. Mental ages have been compared with chronological ages by dividing the mental age of a pupil by his chronological age.³ The resulting quotient is called the intelligence

¹ Monroe, Walter S. *The Illinois Examination*, Bulletin No. 6, Bureau of Educational Research, University of Illinois Bulletin, vol. xx, No. 9.

² It is likely that this statement is true for most other fields of ability.

³ More strictly speaking, a pupil's mental age is divided by the norm for his chronological age.

quotient, or I.Q. McCall¹ and Franzen² have divided the pupil's achievement age by his chronological age and called the result the pupil's educational quotient (E.Q.). This quotient indicates the relation of a pupil's achievement to his chronological age, or, more strictly speaking, to the norm for his chronological age. If a pupil's achievement age is just equal to the norm for his chronological age his E.Q. will be 100. If his achievement age is above the norm for his chronological age his E.Q. will be greater than 100. If it is below it will be less than 100. The magnitude of a pupil's E.Q. will, therefore, define the position of his achievement with reference to the norm for his chronological age.

The comparison of a pupil's achievement with the norm for his chronological age is not in itself very valuable for most purposes. It does not tell us whether the pupil's achievement is what it should be, because we know that pupils of the same chronological age differ widely with respect to capacity to learn. In order to secure a comparison of a pupil's achievement with his capacity to learn, as defined by his mental age, Franzen has proposed that a pupil's E.Q. be divided by his I.Q.

$$\frac{E.Q.}{I.Q.} = \frac{\frac{A.A.}{C.A.}}{\frac{M.A.}{C.A.}} = \frac{A.A.}{M.A.}$$

This, however, is essentially the quotient of a pupil's achievement age divided by his mental age. Thus, a pupil's achievement is compared with his capacity to learn, or, more strictly speaking, with the norm for pupils possessing his capacity

¹ McCall, W. A. "A Proposed Uniform Method of Scale Construction"; in *Teachers College Record*, vol. xxii, p. 31 (January, 1921).

² Franzen, Raymond. "The Accomplishment Quotient"; in *Teachers College Record*, vol. xxi, p. 432 (November, 1920).

to learn. Franzen has called this quotient the accomplishment quotient (A.Q.).

The writer¹ has proposed, for silent reading and arithmetic, that a pupil's achievement age be divided by his mental age, or more strictly by the norm for his mental age, and has called the quotient an achievement quotient (A.Q.).² Although the plan gives results which are numerically identical with the accomplishment quotient secured by Franzen by a less direct method, there is one significant logical difference. The norms used by the writer in securing achievement ages were mental-age norms and not chronological-age norms. Since this translation into achievement age was made primarily to facilitate comparison with mental age, it logically follows that mental age rather than chronological age should be used as a basis for grouping the pupils in calculating the achievement ages. From the standpoint of the construction of a test Franzen's procedure is more simple, because pupils can be grouped more easily on the basis of chronological age. In order to secure a mental-age grouping it is necessary to give a general intelligence test as well as the achievement test. Since Franzen's method gives the same numerical results, test-makers will probably find it preferable.

The significance of the achievement quotient (A.Q.). The educational quotient (E.Q.) does not appear to possess a great deal of significance as an index of the pupil's effectiveness as a learner. The achievement quotient, on the other hand, is an index of the ratio of the pupil's achievement to his capacity to achieve. The value of the achievement quo-

¹ Monroe, Walter S. *The Illinois Examination*, Bulletin No. 6, Bureau of Educational Research, University of Illinois Bulletin, vol. xx, No. 9.

² A.Q. is an abbreviation for both "accomplishment quotient" and "achievement quotient." Both of these terms have been given essentially the same meaning but the latter appears to be the preferable term. Its use is recommended.

tient will be illustrated in Chapter X, but we may indicate here a general basis for its interpretation. Intelligence quotients (I.Q.'s) are interpreted with reference to fixed norms for pupils of all ages. Pupils who have I.Q.'s of approximately 100 are considered normal or average in intelligence. A pupil who has an I.Q. below 75 to 80 is described as dull, while one who has an I.Q. above 115 to 120 is considered bright. Since the median A.Q. tends to be the same for successive grades, we may proceed on a similar basis to set up levels of interpretation. For Monroe's Standardized Silent Reading Test, Revised, and Monroe's General Survey Scale in Arithmetic, the following basis for interpreting A.Q.'s has been worked out:

<i>Quality of pupil's achievement</i>	<i>Achievement Quotient</i>	<i>Per cent of pupils included</i>
Very superior.....	{ 165 and above }	1
Superior.....	{ 135-164 }	6
Average.....	117-134	13
Poor.....	83-116	60
	71-82	13
Failure.....	{ 55-70 }	6
	{ Below 55 }	1

QUESTIONS AND TOPICS FOR INVESTIGATION

1. In what ways are point scores unsatisfactory as descriptions of pupils' performances?
2. What is a derived measure?
3. Explain McCall's proposed procedure for securing derived measures in terms of scores.
4. What is the "educational quotient"?
5. What is the "achievement quotient"? Compare the achievement quotient proposed by Monroe with the accomplishment quotient proposed by Franzen.
6. Why are percentile scores unsatisfactory?
7. What are the objections to using grade norms as a basis for calculating derived scores?
8. Can the achievement quotient be extended to all school subjects?

SELECTED REFERENCES

BUCKINGHAM, B. R. "Suggestions for Procedures Following a Testing Program: I, Reclassification"; in *Journal of Educational Research*, vol. II, 787 (December, 1920).

FRANZEN, RAYMOND. "Accomplishment Quotient"; in *Teachers College Record*, vol. XXI, 432-40 (November, 1920).

MCCALL, W. A. "A Proposed Uniform Method of Scale Construction"; in *Teachers College Record*, vol. XXII, pp. 31-52 (January, 1921).

MONROE, WALTER S. *Illinois Examination*. Bulletin No. 6 of the Bureau of Educational Research, University of Illinois Bulletin, vol. XX, No. 9. See also *Teachers' Handbook for Illinois Examination*.

PINTNER, RUDOLPH, and MARSHALL, HELEN. "A Combined Mental-Educational Survey"; in *Journal of Educational Psychology*, vol. XII, pp. 32-34 and 82-92 (January and February, 1921).

CHAPTER VIII

NORMS FOR EDUCATIONAL TESTS AND THEIR USES

Definition of norms. As usually calculated, a norm is the median or average of the present attainments of a given group; for example, the grade norms established for Monroe's Standardized Silent Reading Tests are the medians of the scores obtained by giving the test to several thousand pupils in each grade. The norms obtained in this way are used in interpreting the scores of individual pupils, and the median or average scores of classes and other groups of pupils. These norms are commonly thought of as defining the degree of ability which a pupil or a group of pupils *should possess*. In some instances the idea of perfection is associated with a norm. This method of standardizing a test makes it clear that we are not justified in thinking of norms as a statement of what ought to be. As they are now derived, our norms must be interpreted merely as representing the average achievement of groups of pupils under present school conditions.

Types of norms. For the purpose of deriving norms, pupils may be grouped in a number of different ways. The most common plan is to group them according to the school grade to which they belong. The resulting norms are called *grade norms*. The pupils may also be grouped according to their chronological ages. When this is done, we secure *chronological-age norms*. When pupils are grouped according to their mental ages we secure *mental-age norms*. We may also determine norms based upon success or failure in the pursuit of school subjects or vocations. In deriving

success norms, pupils, or adults in the case of vocations, are grouped according to success or failure. The uses of these types of norms will be discussed later in this chapter. Certain general limitations may be mentioned here.

General limitations of norms. The grade placement of pupils is not uniform. Some school systems have only seven grades below high school; others have eight grades; and a few, nine grades. Furthermore, there are marked differences in the promotion rates in different school systems. Consequently, the general plan of grouping pupils according to school grade is variable and depends upon the organization of a school system. This is one reason why grade norms are not satisfactory. Chronological age furnishes the most precise basis for grouping pupils. It is, however, inconvenient to effect such a grouping in standardizing a test because the pupils of a given age are to be found in a number of different school grades. In the case of ages above fourteen, all the children are not to be found in school. A grouping on the basis of mental age requires that all pupils be given a general intelligence test in order to ascertain their mental ages. This greatly adds to the labor of standardizing a test, and there is the added inconvenience that the pupils belonging to any mental-age group must be sought in a number of different grades. The merits of each type of norm, and their uses, will be considered later in this chapter.

The basis of satisfactory norms. As pointed out above, the norms which we now have are merely statements of the average of present conditions. In order that a norm may represent the degree of achievement which a pupil should possess, certain criteria must be satisfied. These may be stated under three headings.

1. **Reasonable expenditure of time and effort.** Investigation has revealed that when subjected to intensive training typical pupils can attain very high degrees of ability.

However, after a certain level is reached, the law of diminishing returns begins to operate, and, unless the ability engendered is very important, the expenditure of a large amount of time and effort cannot be justified on the basis of accepted educational objectives. Therefore, satisfactory norms must be below the maximum of possible achievement. The achievement which is possible for a pupil depends upon his capacity to learn, or general intelligence. What is a reasonable achievement for one pupil may not be reasonable for another pupil belonging to the same school grade. Norms which represent the average of present achievements of pupils are certain not to be too high from the standpoint of being capable of attainment by the "average" pupil with the expenditure of a reasonable amount of time and effort. Our present plan of arriving at norms, therefore, possesses the merit of preventing unreasonable ones for "average" pupils, but it must be supplemented by a recognition of mental age in order to obtain norms for individual pupils.

Investigation has revealed that in many instances the present achievements of pupils are far below the achievements that can be secured by the expenditure of a reasonable amount of time and effort. Therefore, it is not at all certain that the average of present achievements furnishes us with norms that are sufficiently high. It may be that when we learn to adjust our methods and devices of instruction to the capacities of pupils, higher achievements may be obtained with even less expenditure of time and teaching effort.

On the other hand, it is possible that the average of present achievements gives us norms that are too high. This is likely to be the case if our present courses of study provide for greater teaching emphasis upon certain topics than can be justified on the basis of future requirements for the functioning of the ability. Thus the average of present achieve-

ments cannot be accepted as a satisfactory norm unless we can show that our course of study is satisfactory. The standardization of educational tests is thus intimately connected with the determination of minimum essentials and curriculum making.

Within the limits of reasonableness the exact magnitude of norms is determined by the requirements of adult activities and by the requirements of future school activities. A norm must not be thought of as representing the maximum of possible attainments of pupils; in fact, the most defensible norm may be materially below the maximum.

2. Requirements of adult activities. The function of the school is to equip pupils for effective participation in the activities of adult life. Outside of school there are a number of demands which are common to all, or at least to a very large per cent of our population. These include the reading of newspapers, magazines and books, writing letters, expressing ideas in conversation, and the solving of simple arithmetical problems of everyday life. In addition to these, there are a large number of special demands which depend upon one's occupation. In establishing our norms we must take into account the common demands of adult activities for the abilities which the school engenders. Educators differ concerning the extent to which public schools should prepare for the special demands. They probably should receive some consideration in establishing norms, but they should not be a determining factor except in the establishing of norms for special vocational courses. There is need for analyzing the adult activities which are common to a large per cent of our population in order to ascertain the precise demands which they make for the abilities in reading, arithmetic, history, science, etc. Our norms should be somewhat greater than these demands in order to provide for the loss of ability due to infrequent use after children leave school.

3. **Future school requirements.** A satisfactory norm must be efficient with respect to the future instructional requirements of the school. The subjects taught in the school may be classified under two heads: — tool subjects and content subjects. The tool subjects of the elementary school are reading, handwriting, the operations of arithmetic, spelling, and language. In the study of a content subject, such as problems of arithmetic, literature, geography, history, science, etc., the tool subjects are used. Time and effort are required for acquiring skills in these tool subjects. Time and effort are also required when these skills are used in the study of content subjects. If only a small degree of skill is acquired, the time and effort required in the study of content subjects will be greatly increased. For example, time and effort are required in learning to read. Reading is a tool that is used in the study of history. If a pupil enters upon the study of history possessing only a small degree of skill in reading, he will find that much time and effort are required in the study of history. On the other hand, if he is equipped with a high degree of skill in reading he will find that the study of history requires much less time. The work of the school cannot attain its highest efficiency unless the pupils are well equipped with the tools of learning which they have frequent occasion to use in the study of content subjects.

In the seventh and eighth grades and in the high school our manner of carrying on school work by the use of textbooks and reference libraries makes very heavy demands for reading. It is also our custom to require much written work, prepared outside of the recitation period, and in some subjects much written work during the recitation period. This custom makes heavy demands for writing, spelling, and written expression. In arithmetic we expect pupils to learn to solve problems (not examples) by solving problems. In

fact, we require them to solve many problems, and the solving of problems requires the performing of arithmetical operations. In view of the fact that the school makes enormous demands upon its pupils for the use of these tool skills, it is folly not to prepare them adequately for these demands. In case the school fails to equip the pupils to read effectively it means that the numerous assignments which they will be asked to study will not only consume an enormous amount of time, but will also destroy interest in the school work because for them reading is slow and difficult and, hence, a disagreeable task.

School demands often exceed those of adult life. In general discussions concerning what the school should accomplish, and in practically all the discussions of norms for particular school subjects, attention has been focused upon the demands of life outside of school, and the demands of the school have been overlooked to a large extent. This is probably due to the emphasis upon the fact that the function of the school is to prepare children for effective participation in adult activities. This is a most wholesome and commendable point of view, but its acceptance should not blind one to the fact that the demands which the activities of the school make for the use of the skills engendered in the teaching of the tool subjects exceed many of the demands which are made by adult activities that are common to a large per cent of our population. The average man or woman does not meet as pressing demands for reading as do the pupils in the high school. Likewise, the demands for writing, and probably for the other tool subjects as well, which the pupils meet in school are greater than they will meet outside of school.

Even if the tool subjects were not practical, it would be necessary for the school to teach them and teach them well in the first six grades in order that the pupils might be effi-

cient in doing the work of the following grades. It should, however, be recognized that, in the case of the content subjects, the source of the norms is primarily the demands of life outside of school.

Norms stated with reference to specific testing conditions. A given set of norms applies only to certain testing conditions. All elements of the testing conditions which affect a pupil's performance must be specified with reference to a set of norms, and these norms can be used to interpret only those scores which have been secured in accord with these specifications. A similar statement may be made with reference to the rules for scoring the performances of pupils. The explanation of the nature of the test and the directions in regard to methods of work materially affect performances. Hence, in using a test it is necessary to follow explicitly the directions for using it on which the available norms are based.

The effect of acquaintance with a test. The acquaintance of pupils with the testing procedure and with the general character of the test affects their performances on most tests. Pupils who are taking a test for the first time do not earn as high scores as they will make on a second or third trial of the test when it is repeated after only a short interval, even though a different form of the test is used. Consequently, norms which are stated with reference to first-trial scores cannot be used in interpreting scores derived from second or third trials unless these scores are properly discounted.

The effect of acquaintance with the test upon second-trial scores may be determined by having the test repeated soon after the first trial. In a study by the author, the Illinois General Intelligence Scale showed an average practice effect of approximately six months of mental age when the returns from the eighth-grade pupils were not used. In this grade unusual conditions appear to have prevailed, and

when the scores from it were included the practice effect was approximately eight months. For Monroe's General Survey Scale in Arithmetic the average practice effect was found to be approximately 3.2 points in grades three to five, and 4.5 points in grades six to eight. Although the effect of practice has not been generally determined for educational tests, these determinations are probably not larger than what would be found for most of our tests. It appears that in many cases second-trial scores will be about 10 per cent higher than first-trial scores.

The increase of third-trial scores over second-trial scores is less, but not negligible. If an interval of several weeks intervenes between two trials the increase will be less, if no attention has been given to the exercises of the test in the interim. If the exercises have been made the basis of instruction the increase is likely to be greater. The recency of instruction on the topics covered by the test also affects a pupil's performance. Norms which are stated for pupils who have had no recent instruction on the topics covered by a test will not be appropriate for interpreting the scores derived from pupils who have had recent instruction on that topic.

The effect of coaching. When pupils are coached upon the test between trials, or are given training in doing the exercises of the test, much greater gains are to be expected. The Illinois Examination was given in one school where it appears that the teachers decided that their pupils would profit by receiving instruction upon the type of exercise that this battery of tests contains. The teachers did not know that the second form was to be given and, therefore, did not have in mind preparing their pupils for it. Their instruction was to a slight extent based upon the Form I test papers. Finding that their pupils were rather weak in knowledge of vocabulary and in synonym-antonym, special training was given along these lines in language work. In

arithmetic, practice was given upon those combinations in which the pupils seemed weak. In reading there was a special drill for increasing the rate of silent reading.

The first form of the Illinois Examination was given to these pupils in November, and the examination was repeated, using the second form, in May. Although it is commonly assumed that general intelligence is unaffected by school instruction, the difference between the median scores for these two trials was equivalent to slightly more than four years of mental age. Obviously, a median growth of four years in general intelligence is impossible during a period of six months. Thus, we are forced to the conclusion that the training that the school gave these pupils upon the exercises of the Illinois General Intelligence Scale introduced a fictitious gain in the second-trial scores of approximately three and one half years.

One would naturally expect that the achievement scores in arithmetic and silent reading would be materially affected by instruction. The gains were relatively somewhat larger than the gains in mental age. It is probable that some of the increase in the May scores over the November scores was due to the pupils being more familiar with the testing procedure. The explicit training which they received is also a factor. However, some of the gain doubtless represents real growth in achievement, but we have no way of knowing the amount of this growth until we have determined the amount of gain due to coaching and to acquaintance with the testing procedure.

The equivalence of the duplicate forms. Most of our educational tests are being constructed so that there are two or more duplicate forms. These consist of exercises that are the same in kind, but require different answers. The intention is to secure duplicate instruments which will yield equivalent measures. Frequently this ambition is not realized,

Investigation has revealed that duplicate forms may not yield equivalent measures, even though a great deal of care was exercised in their construction to make them equivalent. For example, the three forms of Monroe's Standardized Silent Reading Tests are not equivalent, and it has been necessary to announce correction numbers in order to make comparisons between scores yielded by the duplicate forms.¹ It has also been found that the duplicate forms of the Burgess Picture Supplement Scale for measuring reading ability do not yield equivalent measures.

The equivalence of duplicate forms may be determined by arranging copies of the several forms of a test in alternate order. For example, if it is desired to determine the equivalence of three forms of a test, a pile of test papers should be prepared in which the first paper is a copy of form 1, the second a copy of form 2, the third a copy of form 3, the fourth a copy of form 1, the fifth a copy of form 2, the sixth a copy of form 3, and so on. If this pile of test papers is distributed to pupils as they are seated, each form of the test will be given to a random sample of the pupil population. A sufficient number of pupils should be tested so that two or three hundred scores will be obtained for each form of the test. A comparison of the average scores will indicate the amount of non-equivalence of the different forms of the test, and will also furnish a basis for calculating correction numbers that may be used when it is desired to compare the scores yielded by different forms.²

¹ Monroe, Walter S. *Report of the Division of Educational Tests for 1919-20*. Bulletin No. 5, Bureau of Educational Research, University of Illinois Bulletin, vol. xviii, no. 21, p. 19.

² Monroe, Walter S. *The Illinois Examination*. Bulletin No. 6, Bureau of Educational Research; University of Illinois Bulletin, vol. xix, No. 9. This bulletin gives an account of the determination of the amount of non-equivalence of the different forms of this battery of tests, and also the correction numbers which are proposed for use.

Since duplicate forms seldom yield measures which are equivalent, it is obviously necessary that we have separate norms for each form, or that we use correction numbers to reduce the scores yielded by the different forms to a common basis. Failure to do this may introduce serious errors of interpretation.

Errors of interpretation vs. errors of measurement. Errors of interpretation should be carefully distinguished from errors of measurement. There is no reason to believe that second-trial scores are not just as accurate absolute measures of ability as first-trial scores. In fact, they are probably more accurate, since the pupils are acquainted with the test and hence it is likely that the testings are more nearly constant for all pupils. However, a serious error of interpretation is likely to be introduced if the scores from the two trials are interpreted with reference to the same norms. Errors of interpretation are also likely to be introduced when the directions for administering the test are modified in any way.

Different uses of educational tests require different types of norms. In Chapter III we pointed out a number of activities of the school which require the measurement of the abilities of pupils. Different types of norms are used in the interpretation of measures of ability in these activities. We shall consider the types of norms which these uses of educational measurements require.

Promotion and classification of pupils. In grouping pupils for the purpose of instruction the norms required are naturally those which are based upon the grouping desired. For the placement of pupils in grades, grade norms are demanded for both measures of intelligence and achievement. When the pupils belonging to a given grade are classified into fast, average, and slow sections on the basis of their intelligence quotients, the norms required are those based upon the

probability of success in the respective sections. For example, the norm for a fast section should represent the degree of brightness below which success in the fast section is not likely.

Educational and vocational guidance. Advice with reference to future educational and vocational activities should be given largely on the basis of the probable success in pursuing such activities. For this purpose, we need norms which define the degrees of general intelligence and achievement required for probable success in various educational and vocational activities. Such norms have been obtained by measuring the ability of persons now pursuing these activities, both successfully and unsuccessfully. The resulting information is then sorted on the basis of the success of the persons from whom it was obtained. A norm is thus secured below which success is not probable.

Evaluation of school efficiency. The efficiency of a school¹ system or a subdivision of it cannot be evaluated or measured directly by giving tests to the pupils. Measurements of the achievements are not in themselves a measure of the efficiency of the instruction and training which the school is giving to the pupils. It is only through the comparison of such measures with appropriate norms that we are able to make any inference concerning the efficiency of the instruction and training. If the deviations of pupils' scores from the norm are due, either directly or indirectly, to the instruction and training these deviations then become an index or indirect measure of the efficiency of the school. If they are due to other causes, they will not be a true measure of the efficiency of the instruction. It is, therefore, necessary to inquire into the use of norms in indirect measurement, and into the nature of the various types of norms and their relation to achievement and instruction.

¹ The phrase "efficiency of a school" is used in this chapter in a restricted sense. No account is taken of the educational investment.

The use of norms illustrated : two variables. One of the simplest illustrations of the use of norms is that of the standard, or normal, temperature of the human body. Here, there are just two variable quantities. The temperature of the human body, which may be considered the dependent variable, and the general physical condition, the independent variable. The state of a person's health is very closely related to the temperature of his body. As the state of one's health changes there is a corresponding change in one's temperature. The temperature which accompanies perfect health is taken as the norm. Any deviations from this standard temperature indicate the presence of a pathological condition which needs medical attention. In the case of temperatures above the norm, we commonly express the deviations as so many degrees of fever. A person who has three degrees of fever has a temperature which is three degrees above the norm. The deviation of one's temperature from the norm is a measure of his general physical condition, but the measurement of the temperature of the human body without an appropriate norm with which to compare it would furnish no indication of the state of one's health. The existence of a norm thus makes possible the measurement of one's general physical condition, a thing which we are not able to accomplish by direct methods.

Three variables. Norms are also used in situations that involve three variable quantities. One of the quantities is considered the dependent variable, and the other two are treated as independent variables. The dependent variable depends upon the other two for its value. Any change in the magnitude of either of the independent variables produces a corresponding change in the magnitude of the dependent variable. If one of the independent variables is easily measured by the direct application of an appropriate instrument, norms may be stated which define the amount

of the dependent variable which should, or on the average does, accompany certain amounts of this independent variable. The deviation of the dependent variable from these norms measures indirectly the other independent variable.

This type of situation may be illustrated in the field of physical measurements by the weight, volume, and quality of wheat. The weight is the dependent variable. The volume and quality are the independent variables. The weight and volume are easily measured in a direct way. This is not true of the quality. We may observe that the quality is poor and that the grains are shriveled and have a poor color. Or we may observe that another lot of wheat appears to be excellent in quality. This, however, is not quantitative measurement of the quality of the wheat. In the purchase and sale of this commodity the quality is an important consideration. Hence, it is highly desirable to secure a quantitative measure of it. This is accomplished by means of the deviation of the weight of a sample having a given volume from the standard weight for that volume. The norm for wheat is that 2150.42 cubic inches (a bushel by volume) should weigh 60 pounds. A bushel of wheat does weigh this number of pounds if the quality approximates the average. If the weight of a bushel (by volume) of wheat deviates from 60 pounds, the amounts of the deviation is an index of the quality. For example, if a bushel of wheat weighs only 58 pounds its quality is two units below standard. If a bushel of wheat weighs 63 pounds its quality is three units above standard. Hence, by the use of a standard weight for a given volume we are able to secure an objective measure of the quality of wheat.

Norms in more complex situations. In educational measurements the situations in which we need to use norms are even more complex, due to the fact that instead of only two independent variables there are three or more. The difficul-

ties that are introduced by having more than two independent variables may be illustrated in the field of physical measurements of children. The weight may be taken as the dependent variable, and the height, age, and general physical development as the independent variables. If standard weights are determined for various heights, deviations from these norms are composite indices, or measures, of both age and general physical development.

This, though, is unsatisfactory. The age of a child is not subject to control, and furthermore it is easily measured. What we desire is a measure of the general physical development, which is not easily obtained by a direct method. It, therefore, becomes necessary to do one of two things. One solution of the difficulty is to establish a separate set of standard weights based on height for each age. The other possibility is to show that age does not make a significant contribution to the weight of a child which is not already contributed by his height, and hence may be neglected as an independent variable. If this is done, standard weights established for each height would be satisfactory, and any deviation of the weight of a child from the standard weight for his height would be an index or measure of his general physical development.

The type of norms required for evaluating the efficiency of a school. A pupil's ability of which his performance, or more strictly speaking his score, is a measure depends upon his school grade (amount of instruction), general intelligence, and instruction. This last factor includes quality of instruction, course of study, the pupil's attitude toward his work, and other more subtle elements, all of which are subject to partial or complete control by the school. The effectiveness of the instruction is affected by the way in which the school is organized for instructional purposes. There are other factors, such as chronological age, general physical

condition, home environment, sex, race and nationality, which may be listed as contributing to the achievement of the pupil. The effect of these factors upon achievement is not known with much certainty, but it appears that their contributions are relatively slight except in extreme cases. For these reasons we shall omit them in the following discussion. To include them would greatly complicate our analysis without adding anything to the practical value of the conclusions.

Achievement is the dependent variable. School grade, general intelligence, and efficiency of instruction are the independent variables. If norms of achievement are stated with reference to school grade, the deviation of a score from the grade norm is a measure of the combined influences of general intelligence and instruction. Thus, the interpretation of the deviation of a score from the norm is ambiguous because in the case of a pupil who is below standard we cannot know whether this condition is due to his mental age, over which the school has no control, to the quality of the instruction, over which the school has control, or to a combination of the two factors. A similar statement can be made with reference to the deviations of class scores from grade norms.

Relative influence of general intelligence and school grade upon achievement. There is an increase in the average mental age of pupils from grade to grade. In order to study the relative influence of general intelligence and school grade upon the achievement of pupils,¹ the Illinois General Intelligence Scale and the revised form of Monroe's Standardized Silent Reading Tests, were given to about seven thousand school children in sixteen Illinois cities. These two tests were given to all of the pupils in grades III to VIII.

¹ Monroe, Walter S. *The Illinois Examination*. Bulletin No. 6, Bureau of Educational Research, University of Illinois Bulletin, vol. xix, No. 9.

There is in general an increase in the average rate and comprehension from grade to grade for the same mental age, and also from a lower to a higher mental age within the same school grade. The average increase in achievement from grade to grade for pupils of the same mental age was found to be .3 for comprehension, and 5.7 for rate of silent reading. The average increase in achievement corresponding to an increase of one year of mental age for pupils belonging to the same school grade was found to be 1.1 for comprehension and 9.4 for rate. Thus, the average increase for one year of mental growth is over three times as great for comprehension, and nearly two times as great for rate as the corresponding increases from one school grade to the next. Hence, we are justified in saying that on the basis of these facts the mental age of a pupil makes a much greater contribution to his achievement in silent reading than does the school grade in which he is placed. This being the case, our interpretation of achievement scores in silent reading will be very misleading unless the mental age of a pupil or a group of pupils is taken into account.

It must be borne in mind that the above investigation included only silent reading. It is possible that achievement in other school subjects is not related in this way to general intelligence. It is also possible that the repetition of the investigation in the field of silent reading with other tests would not yield similar results. It is, however, very unlikely that the results would be sufficiently different to indicate the opposite conclusion.

The diagnosis of groups of pupils. Deviations of the class score from the grade norm measure the combined effect of general intelligence and instruction. Since mental age has been shown to be a potent factor in determining achievement, the deviations fail to give definite information concerning the effectiveness of the instruction which the class

has received. If the average mental age of the class is conspicuously above the average for grade, their class score may be up to the grade norm but the instruction may be lacking in effectiveness because, when the intellectual equipment is considered, this class should have an average achievement score notably above the grade norm. The corresponding statement can be made for a class whose average mental age is below that for their school grade. Hence, grade norms are not satisfactory for evaluating the efficiency of the instruction given to a class. The use of mental age norms for this purpose will be considered after our discussion of the diagnosis of individual pupils.

The diagnosis of individual pupils. There are large individual differences in achievement within a school grade. This is true even if the pupils belong to a single class. Thus, the deviations of some individual scores from the grade norm will be large. In a number of cases they will be greater than the difference between the norms for two successive grades. These deviations are measures of the composite effect of general intelligence and instruction. Since they are large, the influence of either general intelligence or instruction or both must be large.

Within a given class the pupils have received approximately the same instruction, but it may not have been equally effective in engendering achievement. Pupils differ in their responses to particular kinds of instruction, as well as in other respects. Thus, it appears reasonable that the effectiveness of the instruction makes a substantial contribution to the deviations from the grade norm.

On the other hand, the application of a general intelligence test reveals very large differences in the mental ages of the pupils belonging to a given class. Differences of four or five years are frequently found. Since mental age is defined as capacity to learn, it is obvious that differences in

the mental ages of pupils are influential causes of the deviations of the achievement scores from the grade norms. Thus, the deviation of a pupil's score from the grade norm is a very poor measure of the effectiveness of instruction. It is not too much to state that in the diagnosis of individual pupils grade norms are of little value. Little information concerning the effectiveness of the instruction is given by the deviation of the pupil's score from the grade norm, unless the mental age of the pupil is also considered.

Grade norms for different mental ages. One way of taking account of the mental age of pupils and classes in the use of grade norms would be to establish separate grade norms for the different mental ages that are found within each grade. This plan would give a very complex system of norms. For the interpretation of individual scores there would need to be at least five norms for each grade. The disadvantage of this plan lies in its complexity.

Mental-age norms superior to grade norms. The fact that mental age appears to make a larger contribution to a pupil's achievement than his grade placement suggests the establishment of mental-age norms to take the place of grade norms. Such norms can be derived by grouping the scores of pupils on the basis of mental age, rather than on the basis of school grade. This will involve a greater amount of labor in the derivation of norms, since it will be necessary to give to each pupil a general intelligence test as well as the achievement test. However, the procedure involves no serious difficulty.

The deviation of an achievement score from a mental-age norm will be an index of the composite influence of school grade and instruction. The investigation referred to above indicated that the school grade makes less contribution to a pupil's achievement than his mental age. Hence, mental age-norms are more satisfactory than grade norms in evaluating the efficiency of instruction.

Comparison with mental-age norms by ratio or achievement quotient. Instead of comparing an achievement score with the norm by subtraction, to obtain a deviation, comparison may be made by forming a ratio. The achievement score divided by the norm will give a quotient which is essentially the per cent of the norm which has been achieved. This is called the achievement quotient.¹

The achievement quotient furnishes a simple method of allowing for the effect of the grade placement of the pupil upon his achievement score. An achievement quotient (A.Q.) of 100 indicates that the pupil's achievement is just up to his mental-age norm. If the grade placement of the pupil makes an appreciable contribution to his achievement, we would expect that a pupil of a given mental age in the sixth grade should have a higher achievement quotient than a pupil of the same mental age in the fourth grade, providing the instruction and other factors, except school grade, conditioning achievement were equal in the two cases. If this is found to be true, standard achievement quotients may be set up for each school grade. Such a dual system of norms will be more convenient than a norm for each mental age within each school grade. It has, however, been shown for silent reading and the operations of arithmetic that the standard achievement quotients are approximately constant in the different grades.²

Advantage of mental-age norms. Certain advantages may be enumerated for mental-age norms. First, mental-age norms emphasize the general intelligence of pupils. This is a good thing, since general intelligence makes a large contribution to the pupil's achievement. Second, the achievement quotient which is obtained by comparing a pupil's achievement score with his mental-age standard is a much better index of the effectiveness of the instruction which the

¹ See Chapter VII.

² See page 159.

pupil has received than the deviation of his score from the grade norm. The achievement of each pupil is compared with his own norm rather than with the norm for the group to which he happens to belong. Third, since mental-age norms are essentially individual, rather than group, emphasis is based upon each pupil doing his best. The achievement of the other members of his group has no effect upon his score, or upon the norm with which his score is compared.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. How are the norms for our educational tests derived?
2. How should we interpret norms with reference to educational objectives?
3. Summarize the merits and limitations of the different types of norms.
4. Why is it necessary to control conditions in the administration of educational tests?
5. Explain how certain magnitudes may be measured indirectly through the use of norms.
6. When a test has been standardized, are we justified in considering these norms permanent or will it be necessary to restandardize a test frequently?
7. Should we expect to have norms for each month of the school year? Why?

CHAPTER IX

HOW TO MAKE A CRITICAL STUDY OF AN EDUCATIONAL TEST

The general outline for a critical study of an educational test. The need for a critical study of an educational test arises both in the construction of a new instrument and in the scientific selection of one of the available tests for use. A refined use of a measuring instrument requires that its limitations — i.e., its deviations from the necessary requirements — be known. The principal items in a critical study of an educational test refer to the extent to which the requirements of test construction, stated in Chapter IV, are satisfied.

The outline for the critical study of a test which follows has been prepared to show the questions which should be answered. Some of the questions call for facts; others require inferences from information which, in some cases, is incomplete. The most important question is that of validity, or the constancy of the functional relation between the scores and the abilities specified by the function. Several of the items called for in the second and third divisions of the outline derive their significance from their use in the consideration of validity. For many of our present tests it is not possible to answer several of the questions in the outline because the necessary information is not available. When applying the outline to such tests the study must, necessarily, be limited to the questions for which information can be secured, unless it is possible to supplement it by original research. Following the outline the questions will be explained, and the method of answering them will be described and illustrated. The pages in the text on which the explan-

ations of the main headings of the outline begin are indicated in parentheses.

OUTLINE FOR MAKING A CRITICAL STUDY OF A TEST

I. Facts of title (p. 185).

1. Author.
2. Exact title.
3. Date of first publication or use.
4. Duplicate forms and parts.
5. Used in what grades?

II. Nature of pupil's performance (p. 185).

1. What the pupil does.
2. How the exercises of the test were constructed and selected.
3. Conditions under which performance is given.
 - a. Explanation of test to pupils.
 - b. Directions as to methods of work.
 - c. Time allowance.
 - d. Distractions.
4. In the case of a scaled test, are the exercises equally spaced on the scale of difficulty?

III. Description of pupil's performance (p. 186).

1. How is the score computed? (This is to include rules for scoring, and derived scores.)
 - a. Individual pupil.
 - b. Class or larger group.
2. What dimensions are described separately?
3. What dimensions are described in combination, and what is the nature of the combination?
4. What dimensions are not described in the pupil's score? Are these constant for all pupils?
5. Does zero mean "not any" of the ability measured? Is the unit constant?

IV. Function of the test, or specification of the abilities the test is designed to measure (p. 186).

V. Validity of the test, or the constancy of the functional relation between the scores and the abilities specified by the function (p. 188).

1. Objectivity in describing performances (p. 196).

2. Reliability (p. 201).

- a. Coefficient of reliability (r_{12}) (p. 202).
- b. Index of reliability (r_{1t}) (p. 206).
- c. Probable error of measurement (P.E.M) (p. 207).
- d. Coefficient of correspondence (p. 218).
- e. Overlapping of successive grade groups (p. 219).

3. Discrimination (p. 219).

- a. Does the distribution of measures agree with the normal curve?
- b. Are differences shown between groups which are known to differ in ability?
- c. Into how many groups is a typical class divided?
Is this sufficient to discriminate properly between the members of a class?

4. Comparison with criterion measures (p. 221).

- a. Teachers' marks.
- b. Measures yielded by other tests.
- c. Composite test scores.

5. Inferences concerning validity, based upon the structure of the test and its administration (p. 226).

- a. Do the content of the exercises and the structure of the test appear to be consistent with its function?
- b. Do all pupils have an opportunity to demonstrate their abilities?
- c. To what extent are testing conditions controlled?
- d. Is the variation of abilities, other than those being measured, reduced to a minimum?

VI. Validity of significance (p. 227).

- 1. Do the traits measured have educational significance?
- 2. Are the measurements expressed in terms of significant dimensions?
- 3. Do the norms form appropriate educational objectives?
- 4. To what school activities is the function of the test applicable?

VII. Norms (p. 229).

- 1. Types of norms available.
 - a. Grade.
 - b. Chronological age.
 - c. Mental age.

2. Representative character of scores on which norms are based.
 - a. Number of scores.
 - b. Type of population groups.
 - c. Time of school year for which norms are stated.
3. Effect of acquaintance with form of test upon standards.
4. Equivalence of duplicate forms.

VIII. Practical considerations (p. 230).

1. Time required for giving and scoring.
2. Cost of test materials.

We shall now consider each of these topical headings, in the order given above.

I-IV. TITLE, NATURE, DESCRIPTION, AND FUNCTION

Facts of title. The questions under the first general head require no explanation, since they call for simple facts.

Nature of pupil's performance. In every test the pupil does something: adds columns of figures and writes the sums, writes answers to questions from memory, makes a drawing, writes sentences from dictation, answers questions based upon a text to which he can refer by underscoring a word, completes sentences by writing missing words, crosses out superfluous words or other symbols, etc. The first question asks for a statement of what the pupil does. It is frequently desirable to supplement the description by quoting a sample of the exercises.

The second question inquires into the procedure followed by the test-maker in determining the particular exercises the pupil is asked to do. For example, in spelling, the questions are, "How were the test words selected with reference to both the course of study and the difficulty of the words?" "What were the considerations which determined the manner of presenting the words?" In silent reading, the questions are similar, "What were the considerations which

lead to the selection of the material to be read?" "Is this material of a particular type? If so, why?" "What considerations lead to the choosing of the type of exercises used?" "How were they constructed?" "Were all the exercises that were constructed used? If not, how were the ones used selected?" The particular questions to be answered vary somewhat with the type of test and with the subject-matter field, but these two cases illustrate their general character.

The meaning of the third question is indicated by the sub-questions. The first two of these distinguish between explaining to the pupil what he is to do — add a column of figures, cross out words, write from dictation, etc. — and how he is to do it — rapidly, carefully, check each exercise, etc. Directions to pupils concerning their methods of work may materially affect their performances.¹ The directions to examiners may also provide for controlling distractions. This is particularly important in individual testing.

Description of pupil's performance. The description of a pupil's performance has been discussed in Chapters V, VI, and VII. After a study of these chapters the questions under this division of the outline should not need explanation. The first one calls for facts which are to be derived directly from the test and the directions which accompany it. Occasionally, users of tests have departed from the directions for scoring specified by the author. Any variations of the method for computing the score should be noted.

Function of the test. The statement of the function of a test specifies the abilities that it is designed to measure. These are the abilities with reference to which the validity of the test is to be judged. Any function may be announced for a test, but apparent inconsistencies between the struc-

¹ Thorndike, E. L., and Courtis, S. A. "Correction Formulæ for Addition Tests"; in *Teachers College Record*, vol. XXI, pp. 1-24 (January, 1920).

ture and content of the test and the announced function must be considered in judging its validity. In such a case experimental evidence must be produced to show that the announced function is fulfilled in spite of these inconsistencies. Any test measures something, although it may do so very inaccurately. A test may be found not to have the function announced for it, but to fulfill another function with a high degree of accuracy. Therefore, one of the results of a critical study of a test may be a redefinition of its function.

For many of our existing educational tests the function has been stated only in very general terms. In some cases, no statement is available except that implied in the title of the test. In making a critical study of such tests, a more exact statement of the function should be formulated by adding the implications of the structure and content of the test. For example, the title, *Courtis Standard Research Tests in Arithmetic, Series B*, suggests that the function of this series of tests is to measure arithmetical ability. An examination of the structure and content, together with a knowledge of the nature of ability in the field of arithmetic, indicates that its function is restricted to the measurement of the arithmetical abilities required for performing the four fundamental operations with integers. Abilities in the fields of common fractions, decimal fractions, and problem solving are not included. The title, *Monroe's Standardized Silent Reading Test*, suggests that the function of this test is to measure the ability to read silently. An examination of the test shows that it consists of exercises which require that the pupil answer questions based upon the paragraphs read. The material to be read is not connected, but it consists of separate exercises. These conditions suggest that a more appropriate statement of the function would be "to measure the ability to read silently simple unconnected paragraphs for the purpose of answering questions."

The statement of the function of an educational test should include specifications with reference to:

1. The particular abilities measured.
2. The type of growth measured. (Two types of growth have been recognized. See page 66.)
3. The type of measures the test is designed to yield — general or specific.

V. VALIDITY OF THE TEST

Under the head of validity we inquire into the degree of constancy of the functional relation existing between the scores yielded by the test and the abilities specified as being measured in the statement of its function. A constant relation means that, for any change in score from pupil to pupil, there is a corresponding change in the abilities specified. Since we use common norms for interpreting all scores, we must investigate this constancy of relation with reference to pupils tested by different examiners at different places and different times.

The validity of physical measurements. In studying the validity of the measurements of physical objects, as, for example, those yielded by a yardstick, we have to consider only two questions,¹ (1) the accuracy of the length of the yardstick and its subdivisions into inches and fractions of inches, and (2) the accuracy with which it is used by competent persons. The first question can be answered by comparing the yardstick with a recognized standard instrument. The accuracy of the measures yielded by the instrument may be studied by having a number of competent persons make measurements from the same object. The fluctua-

¹ If the measurements are made with a high degree of precision, it is necessary to consider, also, the constancy of the measuring instrument. For example, temperature, and, in the case of certain materials, humidity, affect the length of a yardstick.

tions of these measurements from the average constitute an index of their accuracy. These fluctuations may be used as a basis for predicting the magnitude of the departures from a constant functional relation between measures and the magnitude of the objects which may be expected in the measures secured by competent persons.¹

Determination of validity of an educational test complex. The determination of the departures from a constant functional relation between scores yielded by an educational test and the abilities specified by its function is more complex. Ability is measured through a performance, and the performance is described by a score. A brief consideration of the nature of the functional relation between the score and ability will assist in understanding the nature of the problem with which we are dealing in a study of the validity of an educational test. The functional relation of performance to ability has been discussed in Chapter IV. A pupil's performance was shown to depend upon a number of factors other than the ability being measured. Two illustrations will indicate the magnitude of the variations from a constant functional relation which occur under testing conditions which appear to be constant to a high degree. It should be noted that both of these illustrations deal only with the departure from a constant functional relation, in the case of a single pupil, when tested at different times. Only by implication do they indicate the variations from pupil to pupil.

Variability of performance in spelling. Ashbaugh² reports a study in which the same spelling test was given to the same pupils three times within fifteen minutes, and under conditions which were highly constant. The resulting

¹ The question concerning what the yardstick measures is not raised. It is assumed that it measures the straight-line distance between two points, and this assumption agrees with our experience.

² Ashbaugh, E. J. "Variability in Spelling"; in *School and Society*, vol. ix, pp. 93-98 (January 18, 1919).

performances were not consistent. Certain pupils would spell a word correctly the first time and miss it the other two. In fact, there were all possible combinations of correct and incorrect spellings of the same word. The following table shows the per cent of pupils who gave variable performances when given three opportunities, during a fifteen-minute interval, to spell each of twenty words.

Per cent of pupils who gave variable performances:

Grade.....	VI	VII	VIII
Iowa City.....	25	22	18
Hibbing.....	22	25	26

Approximately one fourth of the pupils failed to maintain a constant functional relation between performance and spelling ability during this fifteen-minute interval.

Variability of performance in arithmetic. Tests 7 and 9 of Monroe's Diagnostic Tests in Arithmetic were given to about ninety fourth-grade pupils; tests G and K of the Cleveland Survey Arithmetic Tests were given to about eighty sixth-grade pupils; and two tests in the field of common fractions, one in division and the other in addition, were given to about ninety eighth-grade pupils. Test 7 consists of single column addition examples, thirteen figures to the column. Test 9 involves subtraction, the subtrahend in each instance being a three-place number. Test G is multiplication, the multiplier in each case being a single digit. Test K is long division, with a two-figure divisor. In no case does the second figure of the divisor exceed 3. In the addition of fractions it is necessary to find a common denominator which is the product of the denominators of the two fractions. In the division test the fractions are simple, but some reduction is necessary to bring the results to their lowest terms. In the first four tests the exercises are considered to form equal work units, although they are not identical. They are identical in gross structure, and the usual use of

such tests implies the assumption that all of the exercises call for the same abilities. In the last two tests an attempt was made to have the exercises as nearly equivalent in difficulty as possible. In giving these tests, the usual directions were not followed. All pupils, except the very slowest, were given an opportunity to complete their work. Thus, no measure of the rate of work was obtained. Only accuracy was considered.

Variability in accuracy shown by a table. The performance of each pupil was divided into sections of 3, 4, 5, and 6 examples, each, for the purpose of studying the constancy of the accuracy. The amount of change in accuracy, from section to section, indicates one departure from a constant functional relation. These changes in accuracy ranged from an increase of one hundred per cent to a decrease of one hundred per cent. A number of pupils maintained the same accuracy through two sections of their performance. In Table VI, the per cent of pupils who did this is given for the first four sections of each performance. The table is to be read, as follows: On the addition test in grade four, taking three examples in the group, 48 per cent of the pupils maintained the same degree of accuracy in the second group of examples that they exhibited in the first group; 41 per cent maintained the same degree of accuracy in the third group as they exhibited in the first group, and so on. It will be noted that the per cent of pupils who maintained the same degree of accuracy varies with the different tests. Since these tests were given in different grades, some of this change may be due to differences in the maturity of pupils. However, it is likely that it is due, in part, to the operations performed and, in part, to differences in the length of the examples. It will be noted, also, that there is a small decrease in the per cent of pupils who maintained the same degree as the number of examples in the group is increased. This suggests

that there is some relation between the consistency of performance and the length of the performance.

Both illustrations point to the conclusion that the functional relation between performance and ability is not constant for many pupils, even when the testing conditions are

TABLE VI. PER CENT OF PUPILS WHO MAINTAINED THE SAME ACCURACY OF PERFORMANCE THROUGH TWO GROUPS OF EXAMPLES

<i>Number of examples in groups</i>		<i>Grade</i>	<i>II-I</i>	<i>III-I</i>	<i>III-II</i>	<i>IV-I</i>	<i>IV-II</i>	<i>IV-III</i>	<i>Av.</i>
Addition	3	IV	48	41	39	38	38	40	40.7
	4		36	32	31				38.0
	5		31	32	37				33.3
	6		30						30.0
Subtraction	3	IV	49	54	56	51	52	52	52.3
	4		24	41	51	43	50	42	41.9
	5		46	44	43	46	62	39	46.7
	6		41	41	38	45	33	48	41.0
Multiplication	3	VI	48	46	56	41	52	43	47.5
	4		41	38	37	32	26	45	36.5
	5		31	37	35	39	29	35	34.3
	6		27	25	33				28.3
Division	3	VI	63	65	77	72	67	75	69.8
	4		65	67	72	69	66	77	69.3
	5		62	63	67	64	77	67	66.7
	6		60	66	69	62	58	58	62.1
Addition of fractions	3	VIII	73	71	77	64	78	70	72.1
	4		69	63	67	61	70	67	66.1
	5		68	55	58	44	52	60	56.1
	6		57	43	51				50.3
Division of fractions	3	VIII	65	57	69	55	64	61	61.8
	4		60	49	59	56	56	47	54.5
	5		54	50	56	51	52	52	52.5
	6		53	60	48				53.7

as near constant as can be made. In the general use of educational tests, the testing conditions are more variable. Different examiners administer the tests, and the tests are given at different times and in different places. Furthermore, when we consider constancy of functional relation from pupil to pupil, instead of comparing each pupil with his own record, we would expect to find still greater variations. It, therefore, appears that the functional relation between performance and ability may be characterized as variable. We have assumed in this discussion that a pupil's ability in a given field was constant throughout a short time interval. It may be contended that it is variable, but in this case we are concerned with his average ability and the above statements apply.

The functional relation of performance to ability. In Chapter IV, page 68, the relation of performance to the factors upon which it depends was represented by the equation:

$$P = f(a_1, a_2, a_3, \dots a_n, x_1, x_2, x_3, \dots x_n)$$

in which P stands for the objective performance; $a_1, a_2, a_3, \dots a_n$ the abilities¹ being measured; and $x_1, x_2, x_3, \dots x_n$ the other factors which affect the performance. The x 's include, as general factors, such things as the effort which the pupil makes, his physiological condition, emotional status, and degree of concentration. Other x factors, such as the writing of the sums in a test on addition combinations, or spelling, vocabulary, and other language abilities in a silent reading test requiring reproduction, are sometimes involved.²

¹ In this connection it is, perhaps, helpful to distinguish between the ability a pupil possesses but may not exercise completely because he makes little effort, and the ability which functions in the production of the performance. The former we may call his *potential ability*. The ability which is active is his *kinetic ability*. This is what we measure.

² It is sometimes possible to define ability so that the element of writing is included in it. In the case of the operations of arithmetic, it seems desir-

The functional relation between score and performance. The score obtained from a given performance depends upon certain factors other than the performance. The unit used, its constancy, the location of the zero point, the method of scoring, the training of the scorer, and the objectivity of the scoring are potent factors in determining the score. If the method of scoring is highly objective, the training of the scorer will affect the scores only slightly; but, in the case of such school subjects as handwriting or English composition, the training of the scorer is a very potent factor. Employing again mathematical symbolism, the relation between the score and performance may be expressed by the equation:

$$S = F(P, y_1, y_2, y_3, \dots y_n)$$

in which $y_1, y_2, y_3, \dots y_n$ include such factors as those just mentioned and human fallibility.

The functional relation between the score and ability. The functional relation existing between the score and the ability can be expressed by the following equation, in which the symbols have the meaning already given to them:

$$S = F[f(a_1, a_2, a_3, \dots a_n, x_1, x_2, x_3, \dots x_n) y_1, y_2, y_3, \dots y_n]$$

This relationship may be expressed more briefly by the equation:

$$S = \phi(a, x, y)$$

In this equation a represents the total of the abilities, and x and y the totals of the two classes of factors.

Methods of determining validity. The most direct method of determining the validity of a given test would be to secure, independently of it, measures of the abilities it is designed to measure which are known to be highly accurate, and then to

able to do this except, perhaps, for the fundamental combinations and other simple types of exercises which are usually performed mentally in the doing of a more complex type of example.

compare the measures yielded by the test in question with these. The lack of agreement would constitute an index of the validity of the test. However, this method is not possible, in general. The difficulty is that we are not able to secure independent measures of the abilities which are recognized as highly accurate. In fact, standardized objective tests are proposed as instruments which will yield more accurate and, hence, more valid measures of mental abilities than we are able to obtain by other means. We are, therefore, compelled to study the validity of the function of most tests by methods which are indirect and open to certain limitations.

In a few instances there are available independent measures which are recognized as having a high degree of validity. For example, in the case of intelligence tests, the Stanford Revision of the Binet Test is considered to yield measures of intelligence which have a high degree of accuracy. An index of the validity of another intelligence test may be obtained by comparing the scores yielded by it with the corresponding Binet scores. A common method of comparison is by correlation; and the coefficient of correlation between the two sets of scores becomes an index of the validity of the test. Some tests, principally tests of general intelligence, have for their function the measurement of a pupil's future success in a given field. Such tests are called *prognostic*. The measures of success in school work are definite, even if they are not always true measures of the pupil's real accomplishments. Since they are definite, they may be used as criteria for determining the validity of the function of such prognostic tests.¹

¹ See Kelley, T. L. "The Reliability of Test Scores"; in *Journal of Educational Research*, vol. III, pp. 370-79 (May, 1921), for a discussion of the need for knowing, also, the reliability of a test because the available criteria are not perfectly reliable.

1. Objectivity in describing performances

The "personal equation" in testing. One source of departure from a constant functional relation between the score and the ability is the users of the measuring instrument. Different examiners may not secure the same performances from the same pupils due to differences in the administration of the test. Since numerical measures of these differences cannot easily be obtained, the conditions under which variations in the administration of the tests are to be expected will be considered under the head of "inferences concerning validity." The description of the performances of pupils offers another opportunity for differences in the scores obtained by different examiners. In such a field as the operations of arithmetic the scoring can be made objective, except for chance errors on the part of scorers. Only one answer can be correct. If credit is to be given only for answers entirely correct, there can be no differences of opinion concerning when to allow credit and when to give no credit. In some other subject-matter fields, exercises have been constructed which permit of only one correct answer. When this is done, the test is objective with reference to the scoring.

When the scorer is asked to exercise judgment concerning the quality of a performance, different scorers will give different scores. This occurs in the use of a quality scale in rating samples of handwriting and in describing reproductions of material read. It also occurs in scoring answers to certain types of questions. The subjectivity of the description of such performances may be illustrated by the measurement of English compositions. W. W. Theisen ¹ reports the rat-

¹ Theisen, W. W. "Improving Teachers' Estimates of Composition Specimens with the Aid of the Trabue Nassau County Scale"; in *School and Society*, vol. VII, pp. 143-50 (February 2, 1918).

ing of a number of compositions by each of fifteen teachers. The Nassau County Supplement to the Hillegas Scale was used as an instrument to assist in describing the quality of these pupil performances. The values assigned to the compositions of this scale range from 0.0, 1.1, 1.9, to 9.0.

TABLE VII. TEACHERS' RATINGS OF COMPOSITIONS USING NASSAU COUNTY SUPPLEMENT TO THE HILLEGAS SCALE
(Arranged from Theisen's Report)

Teacher	Compositions											
	F-3	C-7	B-8	B-4	I-7	C-6	B-5	C-3	B-10	G-3	F-9	A-9
1.....	3.8	8.0	5.0	1.9	7.2	8.0	9.0	6.0	5.0	2.8	5.0	6.0
2.....	0.0	6.0	3.8	2.8	8.0	7.2	7.2	6.0	3.8	2.8	5.0	3.8
3.....	2.8	8.0	7.2	8.0	8.0	8.0	9.0	7.2	5.0	3.8	3.8	5.0
4.....	1.9	8.0	7.2	2.8	8.6	8.0	9.0	7.2	5.0	3.2	5.2	5.0
5.....	2.8	7.2	3.8	2.8	9.0	8.0	9.0	7.2	6.0	3.8	6.0	5.0
6.....	1.9	6.0	3.8	2.8	6.0	8.0	8.0	6.0	5.0	2.8	2.8	3.8
7.....	1.9	7.2	6.0	2.8	8.0	9.0	7.2	3.8	5.0	2.8	5.0	1.9
8.....	0.0	9.0	6.0	2.0	8.2	8.0	9.0	7.3	5.4	3.8	6.0	6.0
9.....	1.9	8.0	6.0	0.7	8.0	9.0	9.0	5.8	5.0	1.2	4.2	1.9
10.....	2.8	7.2	6.0	2.8	8.0	9.0	9.0	7.2	6.0	3.8	2.8	2.8
11.....	3.8	7.2	2.8	2.8	7.8	9.0	7.2	5.0	5.6	3.8	2.8	6.0
12.....	1.9	7.2	7.2	2.8	7.2	8.0	6.0	8.0	3.8	6.0	6.0	2.8
13.....	2.1	9.0	6.0	0.7	7.2	8.0	7.2	9.0	6.0	2.8	5.0	5.6
14.....	1.2	7.2	6.0	3.8	9.0	9.0	8.0	9.0	4.4	3.8	7.2	2.9
15.....	3.8	8.0	7.2	2.8	8.0	9.0	8.0	2.8	3.8	3.8	3.8	5.0
Standard value..	2.0	7.4	3.6	1.3	9.3	8.0	8.3	5.9	4.7	2.5	4.0	3.2

An examination of Table VII¹ clearly shows that these teachers did not agree on the description of the quality of any of the composition included in this study. The descriptions of composition F-3 range from 0 to 3.8. For composition C-3 the range is from 2.8 to 9.0. When it is remembered

¹ Values between the steps of the scale were assigned in some cases unless there are some errors in Theisen's report.

that the total range of quality on this scale (from 0.0 to 9.0) is approximately that exemplified by the writings of elementary school pupils, the differences in the ratings by these teachers become very significant. These ratings were made after the teachers had received only slight training in the use of the scale. A few hours of training will materially increase the objectivity of the descriptions of the quality of compositions.

Constant and variable errors. Subjective scoring involves two types of errors, constant errors and variable errors. A constant error results in scores which, in general, are too high or too low. A liberal attitude toward the performances of pupils will result in high scores. On the other hand, a conservative procedure will result in low scores. An indication of the presence of a constant error may be secured by comparing the averages of the two sets of scores assigned independently by two scorers to the same set of test papers. Any differences in their general policy will be reflected by a difference between these averages. However, the difference cannot be considered to be an exact index of the magnitude of the constant error because both persons may be inclined to be liberal in their scoring or both may be conservative, or one may be conservative and the other liberal.

Variable errors are indicated by the fact that, in scoring one performance, Scorer A will assign a score of 90 and Scorer B a score of 75; but, in scoring a second performance, Scorer A may assign a score of 70 and Scorer B a score of 80. This may happen, although Scorer B is, in general, more liberal than Scorer A. In studying the variable errors it is necessary to isolate them from the constant errors. The latter do not affect the coefficient or correlation; hence, it may be used as an index of the magnitude of the variable errors.

Constant and variable errors involved in scoring repro-

ductions. Table VIII gives data relative to both the constant and variable errors involved in the word-counting method of scoring reproductions. The scorers are represented by letters. The numbers in the column headed "Difference of average scores" were obtained by subtracting the average of the scores assigned by the second scorer from the average of the scores assigned by the first scorer. A

TABLE VIII. SUBJECTIVITY OF SCORING REPRODUCTIONS
BY THE WORD-COUNTING METHOD

<i>Test</i>	<i>Form</i>	<i>Grade</i>	<i>Number of scores</i>	<i>Scorers</i>	<i>Difference of average scores</i>	<i>r₁₂</i>	<i>P.E. Est</i>
Memory.....	I	IV	92	Y-C	- 9.9	.89	4.5
"	I	IV	27	Y-K	- 5.1	.92	3.4
"	II	IV	116	Y-C	- 2.0	.90	3.3
"	I	VII	123	Y-K	- 7.5	.77	5.5
"	II	VII	100	Y-C	- 8.2	.84	3.9
"	II	VII	31	Y-K	+ 4.1	.90	2.6
Reproduction....	I	IV	94	L-K	+ 6.8	.97	3.1
"	II	IV	31	L-C	- 1.6	.98	2.4
"	II	IV	68	L-K	+ 4.7	.92	4.2
"	I	VII	117	M-F	- .5	.96	9.2
"	II	VII	113	F-C	- 6.0	.97	5.5
Brown.....	I	IV	111	T-M _y	+12.8	.83	6.1
"	II	IV	110	T-M _y	+ 6.9	.88	4.8
Starch (No. 7)...	I	VII	119	M-C	- 5.8	.97	2.6
" (No. 6)...	II	VII	121	M-C	- 2.0	.97	2.1

positive difference means that the first scorer gave, on the average, higher scores than the second. A negative difference has the opposite meaning. In some cases the difference closely approximates zero, but in others it is relatively large. This indicates that for some scorers the constant error is relatively large. One is justified in asserting that,

on the basis of the possible constant error in the scores assigned to reproductions by a single scorer, no reliable inferences can be made concerning the differences in reading ability of two groups of pupils unless the differences between their average scores are large. It appears that a scorer is not always consistent with reference to his constant errors. In Table VIII, Scorer Y and Scorer K show negative differences for two sets of papers, and a positive difference for a third set. This reversal of policy may be due, in part, to the differences in the character of the reproductions; but, doubtless, the unstability of subjective judgment is also a factor.

In the column headed " r_{12} " the coefficient of correlation between the two sets of scores is given. In the next column the probable error of estimate of the obtained scores from the corresponding true scores is given. This was calculated by the formula,¹

$$P.E._{Est} = .6745 \sigma \sqrt{1 - r_{12}}$$

This probable error of estimate should be interpreted as a description of the amount of variable error, or departure of the assigned scores from the true score after the constant error has been eliminated. We may, therefore, speak of the probable error of estimate in this case as the *probable variable error of the scoring*. A probable variable error of scoring of 3.4 means that, in general, the variable error for the two scorers from whom the data were obtained is greater than 3.4 for 50 per cent of the scores. It is also true that for 50 per cent of the scores the variable error is less than 3.4. The significance of the magnitude of this measure of the variable error of scoring becomes more apparent when it is compared with the average score. In this table a probable variable error of scoring of 9.2 does not necessarily mean that the magnitude of the errors was relatively larger than for another

¹ See page 214 for an explanation of this formula.

test which has a probable variable error of scoring of 4.8. The difference may be due to the size of the units and the absolute magnitude of the scores yielded by the two tests.

2. Reliability

Obtained scores and true scores. Two applications of the same test, or duplicate forms of it, to the same pupils will yield many pairs of scores which are not equal, even after corrections have been made for practice effect or other constant errors. Any lack of objectivity will tend to produce differences in the resulting measures. There are, however, additional factors which also contribute to these differences. As we have shown, it is characteristic of children to be variable in their performances. Some children are more variable than others, and generally a child will give a more variable performance in the earlier stages of his learning than later. Many of the causes contributing to these variations are subtle. Some can be partially controlled and kept constant by appropriate explanation of the nature of the exercises to the pupils and by detailed directions concerning the method of doing them. The more subtle causes, such as the pupil's attitude toward the test, his emotional state, his physiological condition, and his preceding mental activity, cannot be controlled except very slightly.

A pupil's *true score* is defined as the average of a large number of measurements by the same test, or by duplicate forms of a test, the conditions of measurement remaining constant.¹ This is in accord with our method of obtaining true measurements of physical objects. For example, in measuring the diameter of a spherical ball different results will be obtained, and the *true measure* is the average of a large number of

¹ True scores cannot be obtained because the testing conditions cannot be kept constant in the case of repeated measurement. However, the concept of a true score will be helpful in our consideration of reliability.

measures. A strict interpretation is that the average is the *most likely* true measure. The *reliability* of a test refers to the magnitude of the differences between the *obtained scores* and the *true scores*. These differences are the variable *errors of measurement*.

The methods of determining the reliability of a test are methods of describing the magnitude of these errors. In these statements it is assumed that the obtained scores involve only variable errors. Constant errors may be introduced for certain groups of pupils, due to the variation of certain testing conditions, such as timing. These errors tend to become variable for a large group which involves different examiners and the administration of the test at different times. Practice effect and acquaintance with the test produce constant errors. The following discussion of methods of determining reliability apply only to the variable errors of measurement. Constant errors have been considered in Chapter VIII in our discussion of norms.

Methods of determining the reliability of a test. There are five main methods used in determining the reliability of a test, and these will now be described. They are:

(a) **Coefficient of reliability.** The degree of reliability of a set of scores may be expressed by the coefficient of correlation between them and the scores obtained from a second application of the test, or the application of a duplicate form of the test.¹ This is called the *reliability coefficient* (r_{12}). The Pearson formula ² is recommended,

$$r_{12} = \frac{\Sigma xy}{N \sigma_x \sigma_y}$$

¹ In correlating the two sets of scores it is unnecessary to have them reduced to the same scale. Constant errors do not affect the coefficient of correlation.

² See Chapter XIII, for the explanation of this formula, and the method of calculating the value of r_{12} .

The meaning of the reliability coefficient (r_{12}) is indicated in the following illustration. Forms 1 and 2 of the Courtis Standard Research Tests in Arithmetic, Series B, were given to 81 sixth-grade pupils. The correlation table for the scores yielded by the addition test is shown in Table IX. It is to be read: Four pupils made a score of four on Form 1; on Form 2, two of these had a score of four, one a score of five, and one a score of six. From this table the coefficient of correlation was calculated according to the Pearson formula. This was found to be $r_{12} = .87 \pm .02$. This number indicates the degree of general relation between the number of examples attempted on Form 1 and the number attempted on Form 2 by the various pupils. The table shows some attempted more examples on the second form than on the first. Others attempted more on the first than on the second. A few attempted the same number on both forms. If all pupils had attempted exactly the same number of examples on both forms the coefficient of correlation would have been 1.00. However, they did not do this, and this fact is indicated by the coefficient of correlation being less than 1.00. The amount of difference between .87 and 1.00 indicates, in a general way, the degree of failure of the pupils to make the same score on the two applications of the test. The reliability coefficient merely expresses the degree of correlation which one may expect between any two sets of obtained scores, neither one of which may be considered to be *true* measures.

The range of talent — i.e., the range of the scores of pupils — on which the reliability coefficient is calculated, affects its magnitude. If all the pupils belong to a single grade or half grade it will be smaller than if the pupils were selected from a sequence of two or more grades.¹ For this reason it is nec-

¹ Monroe, W. S. "Effect of Grade Distribution upon Coefficient of Correlation"; in *Journal of Educational Research*, vol. II, p. 776 (November, 1920). This reference gives an extreme illustration of the range of talent.

TABLE IX. CORRELATION TABLE SHOWING THE RELATION BETWEEN THE NUMBER OF EXAMPLES ATTEMPTED ON FORM 1 OF THE ADDITION TEST OF THE COURTIS STANDARD RESEARCH TESTS, SERIES B, AND FORM 2 OF THE SAME TEST. SIXTH GRADE, 81 PUPILS

Form I		Form II																								
0		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Total
1																										1
2																										4
3																										2
4																										6
5																										11
6																										14
7																										9
8																										4
9																										4
10																										2
11																										11
12																										16
13																										9
14																										4
15																										4
16																										2
17																										1
18																										1
19																										1
20																										1
21																										1
22																										2
23																										2
24																										2
Total				1	5	2	15	9	14	6	11	5	1	2	4	1	1	1	1	1	1				1	81

Reliability coefficient (r_{12}) = .87 \pm .02

essary to state the range of talent, and to consider it in interpreting the reliability coefficient. The range of talent is roughly indicated by the range of school grades from which the scores were secured. When the reliability coefficient is calculated from scores obtained from pupils belonging to a single grade its value is, generally, between .60 and .80. In a few instances, reliability coefficients above .90 have been obtained. See Chapter XIII for additional suggestions concerning the interpretation of coefficients of correlation.

Relation of the coefficient of reliability to the length of a test. Brown's ¹ formula expresses the relation between the length of a test and its coefficient of reliability. If r_{12} is the coefficient of reliability of the test and r_n the coefficient of reliability of a test n times as long,

$$r_n = \frac{n r_{12}}{1 + (n-1) r_{12}}$$

In this formula, n may be fractional or less than one, as well as integral and greater than one. Since r_{12} is always less than one, r_n will be greater than r_{12} . Hence, other things remaining equal, increasing the length of a test increases its reliability. By means of Brown's formula it is possible to determine the length of a test necessary to yield any desired coefficient of reliability when the coefficient of reliability is known for a given length of this test. It is only necessary to solve this formula for n ,

$$n = \frac{r_n (1 - r_{12})}{r_{12} (1 - r_n)}$$

To find the required value of n , substitute the given values of r_n and r_{12} in this equation.

Calculation of coefficient of reliability from one application

¹ Brown, Wm. *Essentials of Mental Measurement*, pp. 101-02. Cambridge, 1911.

of a test. When there has been only one application of a test we may secure two independent measures by computing two scores from the alternate exercises of the test. One score may be computed from the performance on the odd numbered exercises, and a second score from the even numbered ones. If we represent the coefficient of correlation between the scores yielded by the half tests by r_h , we may calculate the coefficient of reliability of the whole test by means of the formula,

$$r_{12} = \frac{2 r_h}{1 + r_h}$$

This will be recognized as Brown's formula when $n = 2$. If the length of the half tests is multiplied by 2, we have the original test. Hence, r_{12} is the reliability coefficient of the whole test. This method assumes that the test has been constructed so that these parts constitute approximately equivalent tests. Since the two performances were secured at the same time, the testing conditions are probably more completely controlled than when the scores are secured from two independent applications of the test. Consequently, coefficients of reliability calculated in this way will not possess the same significance as when they are calculated from scores obtained from two independent applications of the test. Furthermore, the method cannot be universally applied.

(b) **The index of reliability.** From the reliability coefficient it is possible to obtain the coefficient of correlation between a given set of *obtained scores* and the corresponding *true scores*. This coefficient of correlation is called the *index of reliability*. If we use r_{12} to represent the coefficient of reliability and r_{1t} to represent the index of reliability, the formula¹ is $r_{1t} = \sqrt{r_{12}}$. Since r_{12} is less than 1, r_{1t} is greater

¹ Kelley, T. L. "A Simplified Method of Using Scaled Data for Pur-

than r_{12} . The index of reliability, r_{11} , is the greatest possible value of the coefficient of reliability, r_{12} .¹ The index of reliability expresses the relation between a set of obtained scores and the corresponding true scores. This is a more useful measure of reliability than the coefficient of reliability. Since the calculation of the index of reliability requires only the extraction of a square root, it should be used in preference to the coefficient of reliability. In interpreting it, the bases of comparison will be correspondingly high. Most indices will be between .75 and .90 instead of between .60 and .80.

(c) Probable error of measurement ($P.E._M$). In the case of an instrument for the measurement of physical objects, as, for example, a vernier calliper, repeated measurements of a given object will differ slightly, which means that most of them differ from the true or average measure. The median of these errors or deviations from the true measure is called the probable error ($P.E.$). It is the amount of error which is exceeded in just fifty per cent of the measurements. With uniform care in the use of the instrument, the $P.E.$ will be approximately the same for all objects of the same kind and of approximately the same magnitude. We are, therefore, able to associate the $P.E.$ with the instrument for the measurement of this class of objects. In doing this it is implied that the instrument is always used under standard conditions. In the measurement of mental abilities a large number of applications of a test are impossible, because the act of measurement changes the ability. Instead of a large number of measurements of the ability of one pupil, two or three measurements of the abilities of each of a large number of

poses of Testing"; in *School and Society*, vol. iv, pp. 34, 71 (July 18 and 25, 1916).

¹ Any calculated r_{11} is subject to a probable error, $P.E.$, and this must be understood in the above statement.

pupils must be used as the basis for calculating the probable error of measurement ($P.E._M$). Two formulæ for calculating this measure of the variable error will be explained.

Calculating the $P.E._M$ of a test: Formula A. When two forms of a test are given to a pupil the obtained scores, S' and S'' , will differ from the true score, T . Thus, $S' = T + e'$ and $S'' = T + e''$, in which S' and e' refer to the score obtained from the application of Form 1 of the test, and S'' and e'' refer to the score obtained from Form 2.¹ The differences are the variable errors of measurement.² Any set of e 's for an unselected group of pupils will form a normal distribution whose average is zero, provided the S 's form a normal distribution. Thus, the e 's are simply deviations from this average. The median deviation from this average is represented by $P.E._e$. When we have two sets of measures for the same group of pupils, obtained from two forms of the same test, we may use $P.E._{e'}$ to represent the median deviation of the errors of the scores obtained from Form 1, and $P.E._{e''}$ to represent the same for Form 2. If the two forms are equivalent, and this is included in our hypothesis, $P.E._{e'} = P.E._{e''} = P.E._M$ of the test.

Let $S' - S'' = D_e$. Some of these differences will be positive and some negative and a few exactly zero. They will, also, form a normal distribution with the average at zero if the original measures are distributed normally.

It is now necessary to assume that the distribution of differences obtained from a group of pupils is equivalent to the distribution of differences which would have been obtained from a large number of pairs of measurements of the same ability of one pupil. This assumption, however, is im-

¹ It is assumed that S' and S'' are expressed in comparable units, and from the same zero point. If they are not they must be reduced to this basis. The method of doing this will be given later. This eliminates constant errors.

² We are dealing here only with the variable errors. The constant errors will be considered later.

plied in our purpose, which is to determine the $P.E._M$ of the test for all pupils, at least for all pupils of a given school grade. Although the actual $P.E._M$ may be different for different pupils and for scores of different magnitudes, as a practical device we want a $P.E._M$ which may be used for the test with all pupils of a given grade even if this $P.E._M$ is only approximate.

If e' is subtracted from e'' the differences will form a normal distribution with the average at zero. The median deviation of this distribution is represented by $P.E._{D_e}$. The relation between the three $P.E.$'s is expressed by this equation: ¹

$$\overline{P.E._{D_e}}^2 = \overline{P.E._{e'}}^2 + \overline{P.E._{e''}}^2 - 2r_{12} P.E._{e'} P.E._{e''}$$

In this equation, r_{12} is the coefficient of correlation between e' and e'' . Since e' and e'' are assumed to be independently subject to chance variations, $r_{12} = 0$. Hence,

$$\overline{P.E._{D_e}}^2 = \overline{P.E._{e'}}^2 + \overline{P.E._{e''}}^2$$

Since $S' = T + e'$ and $S'' = T + e''$, $S' - S'' = e' - e''$, and hence $D_s = D_e$. Thus, the median deviation of the difference of the scores, $P.E._{D_s} = P.E._{D_e}$ because the differences are identical. Thus,

$$\overline{P.E._{D_s}}^2 = \overline{P.E._{e'}}^2 + \overline{P.E._{e''}}^2$$

If $P.E._{e'}$ and $P.E._{e''}$ are known, the $P.E._{D_s}$ can be obtained, but in the present case both $P.E._{e'}$ and $P.E._{e''}$ are unknown and cannot be found directly. $P.E._{D_s}$ is easily found. Thus, remembering that $P.E._{e'} = P.E._{e''} = P.E._M$ of the test,

¹ Thorndike, E. L. *An Introduction to the Theory of Mental and Social Measurements*, pp. 190-93.

Yule, G. Udny. *An Introduction to the Theory of Statistics*, p. 211.

$$\begin{aligned}\overline{P.E._{D_s}}^2 &= 2 \overline{P.E._M}^2 \\ P.E._{D_s} &= P.E._M \sqrt{2} \\ \text{or } P.E._M &= \frac{P.E._{D_s}}{\sqrt{2}}\end{aligned}$$

$P.E._{D_s}$ can be calculated directly from the distribution of differences of the scores, but it is simpler to obtain it from the average of the differences taken without regard to sign. Since the differences are also deviations from their average, zero, and the distribution of the differences is normal, the usual relations between measures of deviation, average deviation ($A.D.$), standard deviation (σ), and median deviation ($P.E.$), hold for the differences of the scores.

Median difference $P.E. = .6745\sigma$.

Average difference $Av.Dif. = .7979\sigma$.

Therefore, $P.E. = .8453 Av. Dif.$ This means, simply, that the amount of difference between the two sets of measures which is exceeded in 50 per cent of the cases is equal to .8453 $Av. Dif.$ This formula makes it easy to obtain the $P.E._{D_s}$, since the $Av. Dif.$ is the sum of the differences, without regard to sign, divided by the number of cases. Substituting for $P.E._{D_s}$ in the above equation,

$$P.E._M = \frac{.8453}{\sqrt{2}} Av. Dif. \text{ of scores.}$$

$$P.E._M = .5978 Av. Dif. \text{ of scores.} \quad (A)$$

Application of method. This method may be illustrated by its application to the scores obtained from 88 fifth-grade pupils when they were given two forms of the addition test of the Curtis Standard Research Tests, Series B. The score of Form 2 was subtracted from that of Form 1. These differences gave the following distribution. The central tendency of this distribution is not 0, but is about

-.6. This is due to the fact that the scores on Form 2 were not reduced to the same scale as Form 2 before the subtraction was made. For this reason, the result is slightly in error, the determination of the $P.E._M$ of the test being slightly too large. A $P.E._M$ of .917 means that half of the scores will be in error not more than .917 of an example.

<i>Difference</i>	<i>Frequency</i>
4	1
3	4
2	4
1	16
0	17
-1	21
-2	12
-3	8
-4	1
-5	3
-6	
-7	1
Total	88

$$Av. Dif. = 1.534.$$

$$P.E._M = .5978 \times 1.534 = .917.$$

Reducing one set of scores to the scale of another set. Unless the two sets of measures are expressed in terms of the same unit and with reference to the same zero point, they are not comparable. It is, therefore, necessary to reduce the scores of one test to the scale of the other, or both to a common scale. A common method¹ is to express each score as a deviation from the average in terms of the standard deviation as a unit. This method also eliminates any constant error between the two sets of scores. The method, however, introduces negative scores, which is undesirable in

¹ Woodworth, R. S. "Combining the Results of Several Tests: A Study in Statistical Method"; in *Psychological Review*, vol. xix, pp. 97-123 (March, 1912).

Kelly, T. L. "Comparable Measures"; in *Journal of Educational Psychology*, vol. v, pp. 589-95 (December, 1914).

computing the *P.E.M.* of a test. They may be avoided by the following modification of the method.

Let S_1 be the obtained score on Form 1, and S_2 the obtained score on the second form of the test, and S'_1 a score equivalent to S_2 but on the scale of the first form of the test. S_1 and S_2 being known, it is desired to find S'_1 . Let A_1 and σ_1 be, respectively, the average and standard deviation of the distribution of the scores obtained from Form 1 of the test. Let A_2 and σ_2 represent the same functions of the distribution of scores obtained from Form 2. Since S'_1 is on the scale of Form 1 it will have the same average and the same standard deviation as S_1 . Expressing both S'_1 and S_2 as a deviation from their respective averages in terms of the standard deviations as units, we have $\frac{S'_1 - A_1}{\sigma_1}$ and $\frac{S_2 - A_2}{\sigma_2}$. Since the two scores S_2 and S'_1 are equal, by definition, when expressed in this form,

$$\begin{aligned}\frac{S'_1 - A_1}{\sigma_1} &= \frac{S_2 - A_2}{\sigma_2} \\ S'_1 \sigma_2 - A_1 \sigma_2 &= S_2 \sigma_1 - A_2 \sigma_1 \\ S'_1 &= S_2 \frac{\sigma_1}{\sigma_2} + (A_1 - A_2 \frac{\sigma_1}{\sigma_2})\end{aligned}$$

This equation gives the relation between the obtained scores on Form 2, (S_2) and the equivalent score (S'_1) on the scale of Form 1. Thus, if it is desired to reduce S'_2 scores to equivalent scores on the scale of S_1 , or Form 1, it can be done by means of this formula.

This equation is really the regression equation of S'_1 on S_2 :

$$S'_1 = S_2 \left(r_{12} \frac{\sigma_1}{\sigma_2} \right) + K$$

The coefficient of correlation, r_{12} , does not appear in the above equation, but the assumption that S'_1 and S_2 are

equivalent scores, but on different scales, implies perfect correlation. Hence, $r_{12} = 1$.

Otis ¹ uses a graphical method for translating the scores on one scale into equivalent scores on another scale. The two sets of scores are plotted with reference to coördinate axes, and a line of relation is drawn through the point represented by the average of the two sets of scores and with the slope of $\frac{A.D._y}{A.D._x}$. Using the same symbolism as above, the equation of the line of relation is

$$S'_1 - A_1 = \frac{A.D._1}{A.D._2} (S_2 - A_2)$$

Since $A.D. = .7979 \sigma$, if a normal distribution is assumed, this equation is equivalent to the equation above. Otis uses this line as a "line relation," and scores on one scale are translated into equivalent scores on the other by taking the other coördinate of a point on the line as the corresponding score on the other scale. Although Otis does this graphically, it amounts to using the formula derived above.

If the two sets of measures are reduced to a common scale by means of the formula

$$S'_1 = S_2 \frac{\sigma_1}{\sigma_2} + (A_1 - A_2 \frac{\sigma_1}{\sigma_2})$$

the averages and standard deviations of the two groups of scores become equal. This means that we have secured the elimination of the effect of those factors which tend either to increase or decrease the average of the scores obtained from the second application of a test, or from an equivalent form, and, also, of those factors which tend either to increase or decrease the standard deviation of the scores. Among the

¹ Otis, A. S. "An Absolute Point Scale for the Group Measurement of Intelligence"; in *Journal of Educational Psychology*, May and June, 1918.

factors which affect the average — i.e., produce constant errors — are the size of the units, the location of the zero point, practice effect, acquaintance with the test, the manner of the examiner, and non-equivalence of the tests. The practice effect and acquaintance with the test probably tend to increase the scores of the bright pupils more than those of the less capable. In so doing, the standard deviation is increased.

Calculating the $P.E._M$ of a test: Formula B. The $P.E._M$ of a test may be obtained from the coefficient of reliability and the standard deviation of the distribution of scores from which the coefficient of reliability is calculated. The probable error of estimate ($P.E._{Est}$) is given by the equation¹

$$P.E._{Est_{12}} = .6745 \sigma_1 \sqrt{1 - r_{12}^2}$$

In this expression, σ_1 is the standard deviation of the distribution of scores which is equal to the standard deviation of the S'' scores, if the two sets of scores are comparable, and r_{12} is the coefficient of correlation between the scores derived from the two forms of the test. The probable error of estimate is a measure of the extent to which Form 1 scores depart from perfect correlation with the corresponding Form 2 scores. We may also obtain a probable error of estimate of the departure of the obtained scores from the corresponding true scores. If we represent this by $P.E._{Est_{1t}}$,

$$P.E._{Est_{1t}} = .6745 \sigma_1 \sqrt{1 - r_{1t}^2}$$

$$\text{Since } r_{1t} = \sqrt{r_{12}}$$

$$P.E._{Est_{1t}} = .6745 \sigma_1 \sqrt{1 - r_{12}}$$

$$P.E._{Est_{1t}} \text{ is the same as } P.E._M$$

$$\text{Hence, } P.E._M = .6745 \sigma_1 \sqrt{1 - r_{12}} \quad (B)$$

¹ See Chapter XIII, for a more detailed explanation of the probable error of estimate.

This formula is simpler to apply than formula A (page 210). Hence, it is recommended for use. If the distributions of scores obtained from the two forms of the test do not yield equal standard deviations ($\sigma_1 = \sigma_2$), it is advisable to calculate both

$$\sigma_1 \sqrt{1 - r_{12}} \text{ and } \sigma_2 \sqrt{1 - r_{12}}$$

or, what is simpler, use the average of σ_1 and σ_2 .

Probable error of measurement expressed as a ratio. The magnitude of the $P.E._M$ of measurement becomes significant only when it is compared with the magnitude of the measures from which it was calculated. A $P.E._M$ of 1.5 units for a score of 100 units represents a much higher degree of reliability than a $P.E._M$ of 1.5 for a score of 10 units. In the former case, the $P.E._M$ is only 1.5 per cent of the score; in the latter, it is 15 per cent of the score. Thus, in interpreting the $P.E._M$ which is obtained for a test it is necessary to compare it with the magnitude of the scores from which it was calculated. A simple method of making

this comparison is by means of the ratio, $\frac{P.E._M}{Av.}$. This

expresses the $P.E._M$ as a per cent of the average score of the group of pupils.

Relation of $P.E._M$ to the length of a test. If the length of a test is multiplied by four, so that, other conditions of the test remaining the same, the absolute magnitude of a pupil's score will be just four times what it was, the $P.E._M$ of the score is affected in its relation to the absolute magnitude of the score in the same way as the $P.E.$ of an average is affected by multiplying the number of cases upon which it depends by four. The formula for the $P.E.$ of an average is

$$P.E._{Av.} = .6745 \frac{\sigma \text{ distribution}}{\sqrt{n}}$$

In this formula, $P.E._{Av.}$ is the median deviation of similar averages from the true average, σ distribution is the standard deviation of the distribution from which the obtained average is calculated, and n is the number of cases from which the obtained average is calculated.¹ Since the absolute magnitude of the standard deviation of the distribution remains approximately constant as n is increased, $P.E._{Av.}$ decreases inversely as \sqrt{n} . If n is multiplied by 4, $P.E._{Av.}$ becomes one half of its value for n .

In the case of the $P.E._M$ of a test which is quadrupled in length,² its $P.E._M = P.E._M$ original test $\times \sqrt{4}$. The ratios

¹ Thorndike, E. L. *An Introduction to the Theory of Mental and Social Measurements*, pp. 188-90.

² This statement is based upon several assumptions. In the first place, a score is assumed to be an average, or, more strictly speaking, a constant times an average. This assumption is not contrary to the observed facts. A pupil's performance may be considered to consist of elements or divisions. For example, in an addition test the perception of each of the number symbols, forming each association, and writing each figure of the sums may be considered elements. The pupil's performance is not constant for all of these elements, either with respect to rate or accuracy. If it were possible to isolate these elements and to describe each separately, the pupil's score would be the sum of these descriptions, which is nothing but the average multiplied by the number of elements. Thus, $P.E._{M_K} = .6745 \frac{\sigma_K}{\sqrt{K}}$

In this equation, K is the number of elements in the test and σ_K is the standard deviation of K elements.

It is now necessary to assume that σ_K is the true σ ; i.e., its magnitude does not change as K is increased, or, perhaps more properly, the changes in its magnitude as K is increased are so small as to be negligible. This assumption is always made in applying the formula

$$P.E._{Av.} = .6745 \frac{\sigma \text{ distribution}}{\sqrt{n}}$$

to averages. Hence, this assumption is in accord with an accepted practice. Upon the basis of these two assumptions we may write

$$P.E._{M_K} = .6745 \frac{\sigma_K}{\sqrt{K}}$$

$$P.E._{M_{nK}} = .6745 \frac{\sigma_{nK}}{\sqrt{nK}} = .6745 \frac{\sigma_K}{\sqrt{nK}}$$

of each $P.E.M$ to the score with which it is associated are

$$\frac{P.E.M \text{ quadrupled test}}{\text{Score of quadrupled test}} \quad \frac{P.E.M \text{ original test}}{\text{Score of original test}}$$

But, by hypothesis,

$$\text{Score of quadrupled test} = 4 \times \text{score of original test.}$$

Thus,

$$\begin{aligned} \frac{P.E.M \text{ quadrupled test}}{\text{Score of quadrupled test}} &= \frac{P.E.M \text{ quadrupled test}}{4 \times \text{score of original test}} \\ &= \frac{P.E.M \text{ of original test} \times \sqrt{4}}{4 \text{ score of original test}} \end{aligned}$$

$$\frac{P.E.M \text{ quadrupled test}}{\text{Score of quadrupled test}} = \frac{1}{2} \frac{P.E.M \text{ of original test}}{\text{Score of original test}}$$

This equation means that by quadrupling the length of the test its accuracy or reliability has doubled; i.e., the ratio of the magnitude of its $P.E.M$ to the magnitude of the scores has been reduced to one half of the original ratio. It must, however, be remembered, this it is assumed, that the original test can be increased in length without producing

$$\text{Dividing one equation by the other, } \frac{P.E.M_K}{P.E.M_{nK}} = \frac{\sqrt{nK} \sigma_K}{\sqrt{K} \sigma_K} = \sqrt{n}$$

$$\text{Whence, } P.E.M_{nK} = \frac{P.E.M_K}{\sqrt{n}}$$

But, instead of using the average of nK elements as the score, we use n times this average when we increase the length of the test n times.

$$P.E.M_{n(nK)} = nP.E.M_{nK}$$

Hence, multiplying the above equation by n ,

$$n \times P.E.M_{nK} = \frac{n \times P.E.M_K}{\sqrt{n}}$$

$$P.E.M_{n(nK)} = \sqrt{n} \times P.E.M_K$$

any change in testing conditions. It is only when this assumption is realized that the above statement is true. In actual practice the assumption will frequently be only partially realized and hence this relation between the $P.E._M$ and the length of a test will be only approximated.

The $P.E.$ of class scores. The discussion up to this point has had to do entirely with the reliability of individual scores. If the average is used as the class score, the formula for the $P.E._M$ of an average applies

$$P.E._{Av.} = .6745 \frac{\sigma \text{ distribution of scores}}{\sqrt{n}}$$

By somewhat the same reasoning as given in the preceding section, it may be shown that

$$P.E._{Av.} = \frac{P.E. \text{ individual score}}{\sqrt{n}}$$

In this formula, n is the number of scores on which the average is based. If the distribution of the scores is normal the median equals the average. Hence, a similar formula may be stated for the median.

(d) **Reliability in terms of the coefficient of correspondence.** The coefficient of correspondence is the per cent of the pupils who maintain the same relative position within the group on the second form of the test that they had on the first form. The method of calculation depends upon the definition given to "maintain the same relative position." One method would be to transmute the scores obtained into ranks. The per cent of pupils who received the same rank in both forms of the test would be the coefficient of correspondence. A more satisfactory method is to express each score as a deviation from its average in terms of some measure of deviation of the distribution of scores. The difference between a pupil's two scores, when expressed in this

form, will be an index of the extent to which he has maintained his "relative position," and we may arbitrarily define "maintaining the same relative position" to mean that this difference is not greater than a certain amount. One writer¹ has arbitrarily defined "maintaining the same relative position" to mean that the difference between a pupil's two scores shall not be greater than $\pm 1.00 A.D.$

(e) **Overlapping of successive grade groups.** It is a well-known fact that groups of pupils from successive school grades exhibit a considerable amount of overlapping in ability. One cause of this overlapping is the unreliability of the measures. In general, the greater the overlapping the greater the unreliability is. Hence, if two tests measuring the same abilities have been given to two groups of pupils known to differ in ability, the amount of overlapping of the resulting scores will afford a rough indication of the relative reliability of the two tests. This method is not recommended for use.

3. *Discrimination*

Certain criteria of discrimination. This index of validity refers to the differentiation of scores for pupils possessing different degrees of ability. A measuring instrument lacks discrimination when it fails to yield different scores for pupils or groups of pupils who possess different degrees of the ability being measured. It is obvious that any lack of objectivity or reliability will result in a lack of discrimination for certain pupils. Therefore, the measures of objectivity and reliability considered in the two previous sections are, indirectly, indices of discrimination. There are, however, certain direct criteria of discrimination which may be applied.

In general, unselected groups of pupils from successive

¹ Courtis, S. A. *Measurement of Classroom Products*, p. 481. Report of Gary School Survey, 1919. -

school grades exhibit an increase in average ability from grade to grade. Therefore, a test which fails to show an increase in average ability from grade to grade, or which shows a very small increase, is likely to be lacking in discrimination. Other things being equal, that test is the best which shows the greatest discrimination between successive grade groups. Similar statements can be made with reference to chronological-age groups, or other groups which are known to differ in the ability measured.

If the distribution of scores obtained from an unselected group of several hundred pupils shows a marked departure from the normal distribution, it is evidence of a lack of discrimination for certain pupils of the group. If a considerable number of pupils make perfect scores we have evidence that, for such pupils, the test lacks discrimination. Similarly, a large number of zero scores is an indication of the lack of discrimination for the pupils making these scores.

Objections to use of a large unit. The use of a large unit which cannot be subdivided will prevent proper discrimination for many pupils. The units into which instruments for the measurement of physical objects are divided are subject to subdivision so that relatively small differences may be expressed, either in terms of the unit or in terms of a fraction of it. Many of our educational tests are of such a nature that fractions of units cannot be used in expressing the scores of individual pupils. For this reason, if the unit used is relatively large, a large number of pupils may receive the same score even though they possess different degrees of the ability being measured. The degree of precision with which measurements of physical objects are made and expressed is not justified in the case of mental measurements, for the reason that we are unable to make mental measurements with the same degree of accuracy. For example, in the case of the *Courtis Standard Research Tests in Arith-*

metic, Series B, it would be absurd, if possible, to express the measure of a pupil's rate of work on the addition test to the hundredth part of a unit. The probable error of measurement for this test is nearly one unit.

Probably the best procedure to follow in studying the size of unit is to ascertain the number of groups into which pupils belonging to the same school grade are divided, on the basis of their scores. If the number of groups is less than ten, when the number of cases is fifty or more, the fact should, in general, be accepted as evidence of a lack of proper discrimination, due to the coarseness of the unit used. On the other hand, one should bear in mind that the division of the scores into ten or more intervals cannot be accepted as evidence of valid discrimination. A test may fail to discriminate with reference to a specified ability because it measures another trait, or because the scores involve errors.

4. Comparison with criterion measures

Teachers' marks. The comparison of scores with teachers' marks or ratings as a method of determining the degree of validity of a test is based upon two assumptions: (1) that the teachers' marks are measures of the same ability that the test measures; and (2) that the marks are more accurate indices of the ability than the scores yielded by the test. Studies of teachers' marks have shown that, usually, they are conspicuously lacking in accuracy, and hence in reliability as general measures. Thus, this method should be used with caution. It is most helpful in the case of an achievement test which is so general that it covers the entire field of a school subject, or in the case of general intelligence tests.

Since the virtue of the method depends upon the reliability of the ratings given by the teacher, care should be taken to secure ratings which will have the highest possible reliability. This is probably best accomplished by some systematic

plan of rating, such as the officers' rating scale which was used in the army. A rating scale of this type for silent reading is given below.¹ This scale was used in a junior high school where there were four teachers who were acquainted with each pupil. The average of the two middle ratings was taken as the best criterion of the pupil's ability to read silently. A rating secured in this way should afford a highly reliable criterion for judging the validity of a general test.

RATING SCALE — SILENT READING

Teacher.....
Grade and Section

Silent reading is defined as ability to comprehend readily and accurately, in reading text-books, magazines, and general literature.

Write on first line (15 value) the name of that pupil, of all you have ever known in the grade you are rating, whom you consider the best reader; similarly, put, on the last line, the poorest reader. Then choose a middle case for the middle line; finally, fill in cases half way between the middle and each extreme.

Now, rate your class as to ability in reading, using the above scale and reference cases. (It is urged that, in rating, the intermediate points, as well as the five reference values given on the scale, be used. The class distribution should, in fact, show more fours than threes, and more fives than fours.)

Scale

15.....
12.....
9.....
6.....
3.....

<i>Pupil number</i>	<i>Name</i>	<i>Rating</i>
1.....		()
2.....		()
21.....		()
22.....		()

¹ Pressey, S. L. and Pressey, L. W. "The Relative Value of Rate and Comprehension Scores in Monroe's Silent Reading Test as Measures of Reading Ability"; in *School and Society*, vol. x, pp. 747-49 (June 19, 1920).

The validity of prognostic tests. In the use of school marks, as a criterion of validity, it is necessary to recognize two types of tests with reference to function. Certain tests are designed to measure the achievements of pupils. These achievements are the results of instruction and training, provided either by the school or by other educational agencies. Other tests are intended to measure a pupil's capacity to achieve within a given field. The function of these tests may be described as prognostic. Achievement tests also fulfill a prognostic function, because the measure of a pupil's success at any time is an indication of his probable future success. Therefore, in a sense, all tests have a prognostic function. We are, however, concerned here with those tests which are specifically designed for prognostic purposes.

A pupil's school marks represent his success. It is true that these marks are not always measures of a pupil's real success, but they do represent his actual success as determined by the school. Therefore, when studying the validity of a prognostic test, comparison with the school marks which a pupil receives in the future constitutes a practical basis for judging the validity of the test. It may be pointed out that the correlation between the scores yielded by a prognostic test and the school marks, which represent the pupil's future success, cannot be expected to be perfect. The failure of some pupils to utilize their capacity to learn, the unequal effectiveness of the instruction which the schools provide and the inaccuracy of teachers' marks all alike combine to reduce the degree of correlation.

Measures yielded by other tests. If two tests are considered to have the same function, and one is known to have a high degree of validity, comparison of the second test with it furnishes an index of the validity of the second test. The method of making this comparison is to have the two tests given to the same pupils under the same conditions, and to

correlate the resulting scores. Except by chance, the coefficient of correlation obtained cannot be greater than the coefficient of reliability of the second test. The value of this method is dependent upon the availability of tests which have a high degree of validity. Since in most fields the validity of our tests has not been determined, this method cannot be said to have much value. The usefulness of this method is also limited by reason of the fact that two tests are seldom considered to have identical functions. Frequently, one reason for the construction of the proposed test is that it fulfills a different function. In general, this method should be considered to yield only a rough indication of the validity of a proposed test. It is, perhaps, most useful in the field of general intelligence, where the Stanford Revision of the Binet Scale for Measuring Intelligence is considered to be highly valid, and has been frequently used as the basis of judging the validity of other tests of general intelligence.

Comparison with composite test scores. This method is based upon the principle that the average of a number of measures is likely to be nearer the true measure than any one of the measures used in obtaining the average. The composite score is the average of the scores of the same pupil after they have been reduced to a common scale. It is based upon scores obtained from different tests which are designed to measure the same ability or group of abilities in the same general field. In considering the validity of a test by this method it is necessary to distinguish between one whose function is general, over a rather large field, and one whose function is limited to a very narrow field. In the case of a general test, the exercises which the pupils are asked to do in one test represent only a relatively small sample selected from the large field. This would be true in spelling. Therefore, the composite score of two or more general tests in the

same field should be a more accurate measure of the pupil's ability than the score derived from a single test, because of the larger sample on which it is based. In case the exercises of a test are different from those which the pupil is accustomed to, different types of performances will, in general, result in composite scores which are more valid than those obtained from a single test. For example, in the case of silent reading, a composite score based upon a group of tests which call for a variety of performances will probably be more valid than scores based upon a single type of performance. One reason for this is that some pupils will excel in one type of reading, and others in another. Since silent reading is a complex activity and we may recognize several different types of reading, a general measure of a pupil's reading ability is the average or a composite of his ability in all types of silent reading.

Comparison with composite measures is particularly useful when the measuring instrument is made up of a battery of sub-tests, as in the case of most group intelligence tests. In the interest of economy it is desirable to reduce the number of sub-tests to a minimum. The procedure usually followed is to try out a large number of sub-tests which, taken together, are considered to cover rather completely the field of general intelligence. The composite of the scores yielded by all of these tests is then formed, and the group of sub-tests which correlate most highly with this composite is selected to form the final measuring instrument. By this procedure the sub-tests are selected which best fulfill the function of a larger group of tests. Therefore, they are considered to yield the most valid measures of intelligence which can be obtained by such a number of sub-tests. In selecting the sub-tests in such cases it is also customary to give attention to the inter-correlation between the various subtests, and select those which exhibit a low degree of correlation with

each other. The reason for this is that when a high degree of inter-correlation is exhibited the two tests tend to fulfill the same function.

5. Inferences concerning validity

These based on structure of test. Under the head of reliability we have considered the accuracy of mental measurements. However, to show that the measurements made by a given instrument possess a high degree of reliability is not sufficient evidence to prove that the instrument measures the ability which it is considered to measure; i.e., that a constant functional relation exists between its scores and the abilities specified by its function. A test always yields some sort of measure of something, but it may happen that it measures an entirely different ability, or a combination of abilities in which the ability it is intended to measure plays a minor part. It is, therefore, necessary to supplement studies of the objectivity, reliability, and discrimination of a test by such inferences as can be made concerning its validity. The basis for these inferences is the structure of the test and the plan of its administration, including the description of the pupil's performance. This is supplemented by the nature of the ability measured. This is the most common method for "sizing up" a test. One examines a test, noting certain characteristics, and forms a judgment concerning its validity. The application of this method is usually not systematic or complete. Significant characteristics may be overlooked. A detail may be singled out and the test condemned or approved on the basis of it alone, whereas a complete consideration of the structure of the test and the nature of the ability being measured would result in the opposite conclusion.

When a test lacks in validity. The characteristics of a test which should be noted, in making inferences concerning

the constancy of the relation existing between the scores it yields and the abilities specified by its function, were considered in Chapter IV, under the head of requirements of test construction. We may summarize them here by stating the conditions which justify the inference that the test is lacking in validity. Our final estimate of the validity of a test should be the consensus of information secured under the head of objectivity, reliability, discrimination, and comparison with criterion measures, plus such inferences as we may be able to make. A test may be considered lacking in validity when any of the following conditions can be shown to exist:

1. The exercises of the test call for the functioning of abilities other than those specified by its function, provided these other abilities cannot be considered approximately constant for all pupils.

2. When a variety of methods of work may be used by the pupils, the method of giving the performance is not specified in the directions to pupils.

3. The existence of testing conditions which do not affect all pupils alike.

4. The abilities specified by the function of the test do not function throughout the test.

5. In the case of a general test, the exercises are not representative of the field of ability defined by the function.

6. Some pupils are not given opportunity to demonstrate their ability.

VI-VIII. VALIDITY OF SIGNIFICANCE, NORMS, AND PRACTICAL CONSIDERATIONS

Validity of significance. The expenditure of time and money which is invested in the application of a test is justified only when the resulting measures are useful in the school activities enumerated in Chapter III. The attainment of a

high degree of validity does not prove the worth of the test unless its function is in agreement with our educational objectives. No matter how accurate or how valid the scores yielded by a test may be, if they are not measures of a trait which it is worth while to have, the measuring instrument must be considered to have little value. For example, a test which measures very accurately a pupil's ability to cross out *t*'s from a printed selection has little value, because the crossing out of *t*'s is a trait which has practically no relation to our educational objectives. In general, the trait which an achievement test measures must be included in our recognized educational objectives if the test is to be considered a valuable measuring instrument. A somewhat similar statement can be made with reference to prognostic tests. In order to be significant, a prognostic test must predict a pupil's success in achieving a worth-while trait.

A test should yield measures of ability in terms of the dimensions which are most significant with respect to our educational objectives. A trait is described in terms of its dimensions of rate, accuracy, and level of difficulty on which it functions. These three dimensions are not equally significant in all fields of ability. In some cases, only one or two are significant. For example, in the case of spelling, the rate is not significant. In solving problems in arithmetic, the rate of work is much less significant than the accuracy of the reasoning process. The level of difficulty is seldom an important dimension; i.e., it is seldom included in our recognized educational objectives. It is desired that pupils acquire abilities, not in order to do difficult things, but in order to do things which are socially useful. It is true that many socially useful things are also difficult, but it is not true that the most difficult things are always the most socially useful. In fact, this is very obviously not true. For example, in the case of spelling, the most difficult words are not the most useful words.

This requirement may be stated from another point of view, in the case of achievement tests. The norms which are obtained are usually interpreted, and rightly so, as objectives or goals to be achieved. Unless these are described in terms of significant dimensions the acceptance of the norms as goals to be attained will be inconsistent with our recognized educational objectives.

Norms. In order to interpret the scores yielded by tests, norms are necessary. In the preceding chapter the different types of norms were discussed, and the limitations of each were pointed out. In the complete description of a test it is necessary to know what types of norms are available. It is also imperative that we know the validity of these norms. One factor to be considered in this connection is the population groups, and the number of cases on which they are based. It goes without saying that the population groups should be representative. Certain population groups appear to differ sufficiently so that separate norms are required. For example, separate norms appear to be required for rural schools and for city schools. If a single general norm is issued to be used with both groups, the scores on which it is based should be obtained from both groups. It is also necessary to know at what time during the school year the scores on which the norms are based were obtained. The directions followed in the giving of the test and in the scoring of the test papers is another factor to be considered. A given set of norms can be used in interpreting only those measures which are obtained by following the same directions.

The evidence at hand indicates that if pupils are acquainted with a form of a test they make higher scores. Therefore, the effect of acquaintance with the form of test needs to be known in connection with the norms. If the norms have been derived from scores which represent a first application of the test, they are not suitable to use in inter-

preting scores obtained from a second application of the test. When a test is used the second time it is desirable to have a duplicate form. For a number of our measuring instruments duplicate forms have been provided. They had been constructed in such a way that they were expected to be equivalent. It has, however, been found that exact equivalence was not secured in most cases. Therefore, it is necessary that the degree of equivalence be known, or separate norms for the different forms established.

Practical considerations. The questions to be investigated under this head need no explanation. They are, however, important questions when a test is being considered for general use. A test which requires a large expenditure of time and whose cost is relatively large may be justified for experimental purposes, but it is not appropriate for general use.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. Why should an outline, such as is given in this chapter, be followed in describing an educational test?
2. What is the value of knowing the reliability of an educational test?
3. Is it reasonable to expect that some time educational tests may be constructed which will yield perfectly reliable scores?
4. What is the meaning of validity of significance?
5. What are the sources of the unreliability of a test?
6. How would you go about increasing the reliability of a test?
7. What are the advantages of expressing the reliability of a test in terms of the probable error of measurement ($P.E._M$)?
8. Why do we experience difficulty in determining the validity of a test?

SELECTED REFERENCES

KELLEY, T. L. "The Reliability of Test Scores"; in *Journal of Educational Research*, vol. III, pp. 370-79 (May, 1921).

MONROE, WALTER S. "The Illinois Examination Bulletin No. 6." Bureau of Educational Research, University of Illinois Bulletin, vol. XIX, No. 9.

MONROE, WALTER S. "A Critical Study of Certain Silent Reading Tests." Bulletin No. 8, Bureau of Educational Research, University of Illinois Bulletin, vol. XIX, No. 22.

OTIS, A. S. "An Absolute Point Scale for the Group Measurement of Intelligence"; in *Journal of Educational Psychology*, vol. ix, pp. 239-61 and 333-48 (May and June, 1918).

OTIS, A. S., and KNOLLIN, H. E. "Reliability of Binet and Pedagogical Scales"; in *Journal of Educational Research*, vol. iv, pp. 121-42 (September, 1921).

CHAPTER X

USING STANDARDIZED EDUCATIONAL TESTS

Standardized educational tests not teaching devices. Occasionally, teachers and supervisors have expressed surprise that pupils did not appear to exhibit any greater degree of ability after a standardized educational test had been given to them than the pupils possessed before. Such persons appear to have expected that in some mysterious way the pupils would be made different by the giving of the test. A standardized educational test is not a teaching device. It is not intended to change a pupil's ability. It is merely an instrument which teachers and other school officials may use to secure information about the achievements of pupils. Whenever such information is needed in carrying on the work in the school, or can be used as a basis for increasing the efficiency of the school, a standardized educational test becomes a means to these ends.

A few persons appear to have thought of standardized educational tests as playthings which might be used for their own entertainment, or for the entertainment of an audience at teachers' meetings. Beautiful charts in colored inks have been used to represent the scores obtained. Because such charts were tangible and new they attracted much attention. Standardized educational tests are, however, not playthings, neither are they teaching devices. It is only as the information which they yield is used in increasing the efficiency of the school that the tests become worth while. We may classify and indicate their usefulness under the two main headings of (I) Teaching, and (II) Supervision.

I. THE USES OF STANDARDIZED EDUCATIONAL TESTS IN TEACHING

In Chapter III we enumerated several activities of the school in which measurements of the abilities of pupils are required. It is not imperative that these measurements be made by means of standardized educational tests, but in many cases more accurate and more definite measurements may be secured by using these instruments. In addition to the uses of standardized educational tests implied in the demands of these activities, these instruments are helpful in certain other respects. They are useful in motivating school work. The norms established for the tests are useful as educational objectives. The results of measurement may also be used in a valuable kind of school publicity.

General considerations concerning the use of standardized educational tests. The use of standardized educational tests involves more than merely giving them, scoring the test papers, and tabulating the resulting scores. Before a test can be given it must be selected. Since we now have available a large number of standardized educational tests, the intelligent selection of an appropriate test is frequently not a simple matter. After the scores are obtained they must be interpreted in terms of the needs of the pupils, and the future school procedure must be planned to meet these needs. If undesirable conditions are revealed they must be remedied. If a high degree of efficiency is shown to exist the work of the school should be directed so that it will be maintained or increased.

1. Selection of standardized educational tests for a specific purpose

Definition of need. In certain of the preceding chapters it has been noted that standardized educational tests may

differ widely with respect to function, even though their titles imply that they have the same function. In Chapter III we indicated the general character of the measures required by certain activities of the school. For certain purposes we need rate tests; for others we need power tests. For some school activities we require general or average measures of the abilities of pupils in a given field; in other cases we need separate measures of specific abilities. For some purposes we require only average or median measures of groups of pupils; for others measures of the abilities of individual pupils are required. In selecting a test one should first define his need. Then, with this need in mind one should carefully inquire into the precise functions of the available tests. Only in this way will it be possible to plan a testing program which will adequately meet one's needs. If the test is to be administered by the teachers it is highly desirable that it be easy to give to the pupils. This is particularly important when standardized educational tests are being used for the first time in a school system. It should be noted that frequently measures of both general intelligence and achievement are necessary.¹

Validity. Standardized educational tests are not perfectly valid; i.e., they do not accurately measure the abilities specified by their function. Some yield measures which involve relatively small errors; others yield measures in which the errors are large. For this reason it is important that one be informed concerning the validity (including reliability) of a test when formulating a testing program. Unfortunately, as we indicated in Chapter IX, there is no single numerical index of the validity of a test. One should, however, ascertain the index of reliability or, better still, the probable error of measurement ² of the tests considered. Unfortunately,

¹ See page 43ff.

² See page 206ff.

this is not always available. When it is not, such information as is available should be secured.

Norms. It is also very important to inquire concerning the nature and validity of the available norms. In Chapter VIII, we pointed out that a given set of norms is valid only for specified testing conditions. Norms which are appropriate for first-trial scores are not appropriate for second-trial scores, unless appropriate corrections are made. Norms which are appropriate for pupils unacquainted with the testing procedure are not appropriate for pupils who are accustomed to taking standardized tests. In case there is more than one form of the test, one needs to inquire concerning the equivalence of these forms.

Cost of testing material. Finally, but not least, one must consider the cost of the testing materials and the time that will be required for their administration. The cheapest test is not necessarily the best, neither is the most expensive necessarily a superior test. In considering the cost of testing materials it must be remembered that frequently a test measures only one ability. In such cases, if we are to measure the achievements of pupils within the field of a school subject, it will be necessary to give a battery of tests rather than a single test. For example, the Courtis Standard Research Tests, Series B, measure ability only in the field of the operations with integers. In order to cover the field of arithmetic it would be necessary to add tests for common and decimal fractions and for the problem field. Furthermore, as we have pointed out in Chapter III, a testing program should include a general intelligence test, as well as tests which measure achievement. This means that frequently a testing program will involve giving not merely one test, but several tests. If only one or two tests are to be given the per-pupil cost will not be large, even if each test is relatively expensive. On the other hand, if the testing program

requires the giving of a number of tests the per-pupil cost will be large unless the testing materials are relatively cheap.

Time cost of testing and scoring. The purchase price of the testing materials is not the only cost. It requires the time of the pupils to take the test. It takes the time of the teacher or someone else to score the test papers, and to assemble the scores that are obtained. Reasonable expenditures of money for testing materials and of time in the giving and scoring of test papers can be abundantly justified. Teachers or other school officials should not look upon educational tests as something additional or superfluous to the work of the school. Certain activities of the school require measurements of the abilities of pupils. If they are not secured by means of standardized educational tests, these measurements must be made in less accurate ways. The use of standardized educational tests should be thought of as an integral and valuable part of the work of the school. The time which a teacher spends in the scoring of test papers or in the giving of a test is likely to be just as profitably spent as the time used in planning a lesson or in holding a recitation. However, the amount of time required should be considered with reference to the amount of information which the test yields. If a test requires the expenditure of a large number of hours of the teacher's time but yields a large amount of information, its time cost per unit of information will be relatively small. On the other hand, a test which requires but little time may be relatively expensive because it yields only very meager information. When the teachers of a school system are just beginning the use of standardized objective tests it is desirable to avoid tests which require a large expenditure of time.

It will seldom be possible to choose a test which is ideal in every respect. It will generally be necessary to effect a compromise. It is imperative that the test selected be appro-

priate to one's needs. When there is opportunity for a further choice, no definite rule can be stated. What is the best test will depend upon circumstances.

2. The administration of standardized educational tests

Testing by a single teacher. When a teacher undertakes a testing program as an individual enterprise he should become familiar with the directions before attempting to administer the test to his pupils. Many of our standardized educational tests are accompanied by sufficiently explicit directions so that a teacher who is willing to invest an hour or two in the study of the test and its directions will have no difficulty in understanding exactly how it is to be administered. In some cases detailed directions to examiners have not been formulated. When this condition exists it will be necessary for the teacher to formulate the details of his procedure. In doing this he should bear in mind that the purpose of giving the test is to secure truthful information concerning the achievements of his pupils. The information will not be truthful unless the testing conditions approximate those for which the norms are stated. Therefore, the teacher should try to approximate as closely as possible the standard testing conditions. Any variation from the standard testing conditions, which tends to either increase or decrease the scores, will tend to make the information untruthful and hence lacking in usefulness.

Administering a testing program throughout a school system. When a testing program is being given throughout a school system it is desirable to have a high degree of uniformity in the giving of the tests and in the scoring of the test papers. Unless the testing conditions approximate uniformity it will not be possible to make valid comparisons between classes or other groups of pupils. Different examiners will tend to vary slightly in the administration of a

test. For this reason it has been urged by some that all of the testing throughout a city should be done by one person, or at most by a small group of persons who have been specifically trained for this purpose. There is no doubt that such a practice will tend to secure a higher degree of uniformity in the administration of a test. However, it is probably best to have the tests administered in each room by the classroom teacher. In the first place the pupils are acquainted with the teacher, and he will be more likely to secure more normal responses than a stranger, especially in the case of younger children. However, the most important reason is that it is extremely desirable to have the teacher in sympathy with the testing program, and for him to be acquainted with the nature of the tests. If the tests are given by anyone else the teacher may feel that they are being imposed upon him from above, and that the results are unjust measures of his work. Unless the tests are given for supervisory purposes the teacher is largely responsible for the interpretation of the measures in terms of the needs of the pupils. He must also plan the details of the remedial procedure and, what is more important, he must execute it. The teacher will be better qualified to do this if he administers the tests and scores the test papers. The loss of uniformity in administration is probably more than compensated for by the gains accruing from the teacher's increased acquaintance with the tests.

It is desirable that a testing program for a school system be in charge of one person, or at the most of a small committee. In cities where there is a bureau or division of educational research the director in charge of the bureau would naturally undertake this work. In other places this may be done by the city superintendent, or by someone designated by him. If the testing is confined to a single building, the building principal would be the logical one to take charge. After the testing program has been formulated the teachers concerned

should be called together. Unless they are familiar with the tests to be given it will be helpful to administer the tests to them in exactly the same way as if they were a group of pupils. This gives the teachers a very definite notion of how to proceed in administering the tests to their pupils. In many cases the tests should be explained to the teachers, particularly their functions. There should be a definite attempt to create an attitude on the part of the teachers that a test is used to secure truthful information concerning the achievements of their pupils. They should be given to understand that truthful information cannot be secured unless there is close conformity with the directions for administering the test, and that this requires explicit and detailed compliance with the directions.

The scoring of test papers. Many of our standardized educational tests are constructed so that only one answer is correct. In scoring such tests the usual rule is to give no credit for answers that are not correct. When a test of this type is accompanied by a list of correct answers the scoring is simple. It is only necessary for the scorer to compare a pupil's answers with the correct answers, and to compute his score. When a pupil is required to give an answer in the form of a phrase or sentence, a great variety of performances will be secured. It is not possible for the test-maker to anticipate all of the answers which pupils will give. Even if such a test is accompanied by elaborate directions for scoring there will still be difficulty in deciding whether some answers should be counted as correct or not. In such cases the scoring is not a simple matter. In order to secure uniformity the one in charge of a testing program should call a meeting, at which a number of representative papers should be taken up and scored by the group as a whole. By doing this supplementary rules for scoring can be formulated. This will tend to increase materially the uniformity of the scoring.

Training in the use of a quality scale. As we pointed out in Chapter VI, the description of a pupil's performance in terms of a quality scale is subjective. Unless the scorers have had appropriate training the scores assigned by them will involve large constant errors as well as large variable errors. A reasonable amount of practice will materially reduce the magnitude of both types of errors. Therefore, in the measurement of ability in handwriting, written composition, or other subjects where a quality scale is used, the teachers should be called together for a period of training in the use of the quality scale. Each teacher should be provided with a copy of the scale, and the general procedure to be followed in using it should be discussed. The teachers should then be asked to rate a number of representative samples.

It is desirable that these be samples of known value. In case such samples are not available, representative samples taken from the pupils' test papers may be used. The samples should be numbered and each teacher should keep a record of the score which he assigns to each paper. These should then be assembled in a summary tabulation.¹ This will probably show marked differences in the scores assigned to the same paper. The teachers who assigned extreme scores probably did so because they gave attention to some characteristic that was not considered important by the majority of the teachers. They should be cautioned to avoid being unduly influenced by such characteristics. The averages of the scores assigned by different teachers will give some indication of the constant error. One will likely find that some teachers in general are inclined to assign scores that are too high, others scores that are too low. If the true values of the samples are known they should be given to the teachers, and each teacher asked to compare

¹ See page 197.

his scores with the true values. In case the true values of the samples are not available, the average of the scores assigned to them by the group may be used as a basis of comparison. A small amount of training will materially decrease the subjectivity of the scoring.¹

Scoring by pupils or clerks. The scoring of test papers, especially in the case of a comprehensive testing program, requires the expenditure of a considerable amount of time. When it is desirable that the test scores be reported promptly to a central office, this places a considerable burden upon the teachers. In order to eliminate the drudgery of scoring it has been proposed that the test papers be scored in the upper grades of the elementary school, by reading the correct answers to the pupils and having them mark the test papers. When no questions can arise concerning what answers are to be accepted as correct, this procedure will generally prove satisfactory. The pupils should exchange papers before the scoring is begun. A rescoring by other pupils will eliminate most of the errors. In some schools the pupils in the seventh and eighth grades have been asked to score in a similar way the test papers for the lower grades. Whenever the scoring is done by pupils, a few papers selected at random should be checked by the teacher or some other competent person.

When questions arise concerning what answers should be accepted as correct, the scoring cannot be done satisfactorily by pupils. In such cases the teacher may be relieved of the burden of scoring by employing competent clerks. In some cases pupils from teacher-training classes in the high school,

¹ Van Wagenen, M. J. "The Accuracy with which English Themes may be Graded with the Use of English Composition Scales"; in *School and Society*, vol. ix, pp. 441-49 (April, 1921).

Theisen, W. W. "Improving Teachers' Estimates of Composition Specimens with the Aid of the Trabue Nassau County Scale"; in *School and Society*, vol. vii, pp. 143-50 (February, 1918).

or from some of the commercial classes, have been drafted for this work. If such pupils are selected with reference to their accuracy in clerical work, and are given preliminary training, the scoring can be done satisfactorily under careful supervision.

It should, however, be noted that a teacher who does not score the test papers of his pupils misses an opportunity for an intimate acquaintance with their mental processes. This is particularly true in the case of those tests where it is necessary for the pupil to formulate his answers in terms of phrases or sentences. Frequently, the form of a pupil's answers will reveal much more concerning his instructional needs than will his score on the test.

3. Planning remedial procedure

Interpretation of scores. The scores yielded by a standardized educational test must be interpreted with reference to the use that is to be made of them. A number of uses will be discussed separately, but there are certain general matters which may be mentioned here. It will generally be helpful to assemble the scores in the form of a distribution. Most standardized educational tests are accompanied by a class record sheet and directions for tabulating the scores. This is usually the first step in their interpretation. From the distribution of scores or from an arrangement of the test papers a class score is calculated. This is, generally, the median, although sometimes the average is used. The median or average score is necessary for interpreting the standing of the class as a whole. It is also helpful in interpreting the scores of individual pupils.

Types of errors. The measures yielded by standardized educational tests involve two types of errors.¹ The variable

¹ See page 198 and 202.

errors of measurement tend to cancel each other in the median or average score of a group. For a group of 25 pupils the probable variable error of measurement of the class score is one fifth as large as it is for the individual pupils. For a group of 100 pupils it is one tenth as large. Many of our standardized educational tests were originally designed to be used in securing average measures of groups of pupils. When used for measuring the abilities of individual pupils the variable errors of measurement are relatively large. In fact, in all measures of mental traits, the variable errors of measurement are much larger than we are accustomed to in the measurement of physical objects. Our best means of describing these errors is in terms of the limit which is exceeded in only a certain per cent (usually 50 per cent) of the scores.¹ This makes it impossible for us to know anything about the actual variable error of measurement in a score of a particular pupil. We can only say what the chances are that it does not exceed a specified limit. For this reason it is necessary to exercise a great deal of caution in interpreting the scores of individual pupils.

It is also necessary to take into account the constant errors. These are produced by some change in the testing conditions or by some other factor which affects all pupils alike. The norms of a test are stated with reference to certain testing conditions. If these conditions are modified in any way, or other factors are introduced which tend to increase or decrease the scores of all pupils, misleading interpretations are likely to be made.

Graphic representation of results. In comparing scores with appropriate norms it is frequently helpful to represent them graphically. This is particularly true when the results of testing are being brought to the attention of a group of teachers or other persons. Sometimes unexpected con-

¹ See page 207.

ditions will be revealed when the scores are represented graphically.¹

Remedial procedure. The scores yielded by standardized educational tests should be interpreted in terms of the remedial procedure which the pupils need. For example, if the scores are being interpreted with reference to the educational guidance of the pupil the interpretation should be stated in terms of the school subjects which he should pursue and should not pursue. This involves more than merely saying that a pupil's score is above, up to, or below the norm. It requires that the one making the interpretation be acquainted not only with the test from which the scores were obtained, but also with school procedure. One cannot interpret scores in terms of remedial instruction unless he is well acquainted with methods of instruction. Neither can one interpret scores with reference to the promotion and classification of pupils unless he is acquainted with this phase of the organization of a school.

The interpretation of scores and the planning of remedial procedure is, perhaps, the most difficult part of the use of standardized educational tests. It is, however, the most important part. Without this step the giving of an educational test will have only limited value. When a test has been given throughout a school system it will be very helpful to call together all teachers concerned and discuss the interpretation of the scores obtained. In fact, the teachers have more need of assistance in this step of the use of standardized educational tests than they have in the administration of them.

The use of standardized educational tests by the supervisor and by the teacher. The following discussion of the

¹ Monroe, Walter S. *Measuring the Results of Teaching*, chaps. III and IX. Houghton Mifflin Company, 1918.

Alexander, Carter. *School Statistics and Publicity*, chap. XI. Silver Burdette and Company, 1918.

uses of standardized educational tests is given in terms of the persons who use them; i.e., the supervisor¹ and teacher. One disadvantage in this procedure is that both the teacher and the supervisor participate in a number of the activities of the school. For the purpose of discussion we have arbitrarily divided the activities into two groups, placing in one those activities in which the teacher is most prominent, and in the other those in which the supervisor is most prominent. However, it will be noted that the teacher has a part in a number of the activities assigned to the supervisor, and the supervisor has part in some of the activities assigned to the teacher.

The teacher's uses of standardized educational tests. The teacher makes use of standardized educational tests in carrying on the following activities:

I. Diagnosis of pupils with respect to achievement for the purpose of planning remedial instruction.

II. Diagnosis of pupils with respect to study procedure for the purpose of assisting them in acquiring good methods of study.

III. Setting immediate educational objectives.

IV. Creating motives for learning.

V. Reporting the achievements of pupils to parents.

In Chapter III we indicated the general type of information which the first two of these activities require. Standardized educational tests are helpful in setting up educational objectives and in creating a motive for learning.

4. Diagnosis of pupil difficulties by the use of standardized tests

Diagnosis with reference to grade norms. Until recently the diagnosis of pupils with respect to achievement con-

¹ The supervisor is used here as a general term to include superintendent, principal, or special supervisor.

sisted merely of comparing the scores with the grade norms, and in noting the character of the distribution of the scores. If the class was being diagnosed as a whole, a class median distinctly below the grade norm was interpreted to mean that the class as a whole needed additional instruction. The pupils might need more of the training that they had received, or possibly their need was for a different type of training. Sometimes this condition was interpreted to mean that the trait shown to be below standard had not received sufficient emphasis. In case the class score was above standard, or the individual scores were too widely scattered, certain inferences were drawn with respect to instructional needs of the pupils.¹

In the case of individual pupils a similar procedure was followed. Scores below the grade norm were interpreted to signify a weakness in the pupil's training. For some reason the pupil had failed to respond in a satisfactory way. He needed additional training of the same kind or of a different kind. Perhaps the pupil had failed to apply himself, and the teacher should take steps to secure a stronger motive.²

Diagnosis of pupils in terms of achievement quotients. The type of diagnosis described above is valuable. A teacher who intelligently uses standardized objective tests in this way will materially increase his efficiency as an instructor. Such diagnosis, however, fails to take account of

¹ Monroe, W. S. *Measuring the Results of Teaching*, chap. v. Houghton Mifflin Company, 1918.

This chapter deals with the interpretation of scores yielded by the Courtis Standard Research Tests in Arithmetic, Series B, according to this procedure.

² Zirbes, Laura. "Diagnostic Measurement as a Basis for Procedure"; in *Elementary School Journal*, vol. xviii, pp. 505-22 (March, 1918).

This is an excellent illustration of the use of standardized objective tests in diagnosing pupils with respect to their achievements.

Burgess, May Ayres. "Classroom Grouping for Silent Reading Drill"; in *Elementary School Journal*, vol. xxii, pp. 269-78 (December, 1921).

the capacity of the pupils to learn. All pupils belonging to a grade are judged with respect to the same norms. We know, however, that all pupils are not alike. Pupils belonging to a class frequently exhibit marked differences in their capacity to learn, although they are in some respects a selected group. Therefore it is not unlikely that two pupils belonging to the same class may make identical scores, but have quite different needs for remedial instruction because they differ in respect to their capacity to learn. For exam-

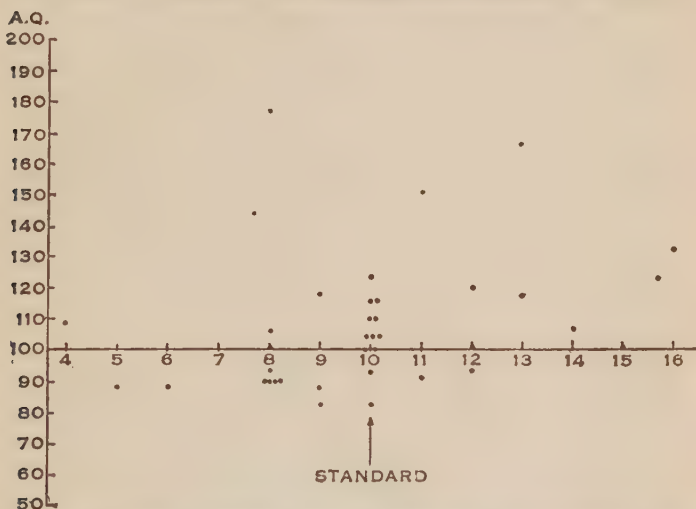


FIG. 5. RELATION BETWEEN POINT SCORES AND ACHIEVEMENT QUOTIENTS IN COMPREHENSION OF SILENT READING, A FIFTH GRADE CLASS.

ple, suppose a "bright" pupil and an "average" pupil make the same score on a test on arithmetic problems. Suppose further that this score is just up to the norm for their grade. The "average" pupil has done as well as we have a right to expect him to do, but the "bright" pupil has not. He should make a much higher score because he possesses a

greater capacity to learn. He needs some sort of remedial instruction in order to bring his achievement up to the norm for his degree of intelligence.

In order to take into account the general intelligence of pupils, mental-age norms have been proposed for use instead of grade norms.¹ The use of such norms makes it possible to compare the achievements of each pupil with the norms for his own mental age. Thus, he is judged with reference to his capacity to learn, rather than with reference to the norm for the grade in which he happens to be placed. It has been proposed that this comparison be made by dividing the pupil's achievement by the norm for his mental age. The quotient obtained is called the achievement quotient.² If this is expressed as a per cent, a quotient of 100 means that the pupil's achievement is just up to the standard for his mental age. If the quotient is less than 100, for example, 80, it means that his achievement is only 80 per cent of his norm. On the other hand, if the achievement quotient is greater than 100, for example, 125, it means that his achievement is 25 per cent above his own norm.

Significance of achievement quotient. The significance of the achievement quotient in diagnosing individual pupils is illustrated in Fig. 5. In this figure the comprehension scores yielded by Monroe's Standardized Silent Reading Test, Revised, are represented graphically on the horizontal scale. The vertical scale is for the purpose of representing the achievement quotients. Each dot in the figure represents a pupil, and its coördinates are his comprehension score on this silent reading test and his achievement quotient. The horizontal position of the dot is determined by the pupil's comprehension score on the test. The vertical position of the dot is determined by his achievement quotient. The grade norm for this class is approximately ten.

¹ See page 179.

² *Ibid.*, 158.

It will be noted that a number of pupils made scores of ten. There is one score of sixteen, and one score of four. An achievement quotient of 100 indicates that the pupil's achievement is up to the norm for his mental age. When the achievement quotient is less than 100 it indicates that his score is below the norm for his mental age. When it is greater than 100 the score is greater than the norm for his mental age.

The highest achievement quotient in this class is earned by a pupil who made a score of eight on the test. He is two units below the grade norm, and would ordinarily be interpreted as being below standard. The norm for his mental age, however, shows that he is not below standard but is decidedly above it. In fact his achievement in silent reading is approximately 80 per cent above the norm for his mental age. It is also interesting to note that the pupil who made the score of four, the lowest in the class, is shown to have an achievement above the norm for his mental age. It is impossible to know what a pupil's instructional needs are until we know what his capacity to learn is. The achievement quotient furnishes a convenient and effective way of recognizing the pupil's mental age in interpreting his achievement.

General diagnosis vs. detailed diagnosis. Achievement within a school subject is complex and consists of many items, and in some cases these are relatively unrelated. A general diagnosis of pupils with respect to achievement involves simply locating pupils who have not achieved general objectives set for them. If we inquire into the particular shortcomings of a pupil we have a detailed diagnosis. This may be extended somewhat indefinitely by the elaboration of our achievement tests. The two types of diagnosis are complimentary, and may be illustrated in the field of arithmetic.

The use of a general test upon the operations of arithmetic which includes examples from each of the four operations may be said to be diagnostic, in the sense of that it will "spot" those pupils who failed to achieve in a general way in the field of the operations of arithmetic. The Courtis Standard Research Test in Arithmetic, Series B, is diagnostic in the sense that it measures achievements separately for each of the operations in the field of integers. It, of course, does not reveal a pupil's particular limitations. It simply tells us the extent to which he is lacking in the general ability to do examples in each of the four operations.

To ascertain the particular types of example which he is unable to do satisfactorily it is necessary to use a more elaborate series of tests, such as Monroe's Diagnostic Tests in Arithmetic, or the Cleveland Survey Tests in Arithmetic. In diagnosing pupils with reference to achievement it is profitable for teachers to use a test, or battery of test, which yield detailed information concerning the achievements of pupils. The time and money at a teacher's disposal, of course, places certain limitations upon the elaborateness of the testing program. In some cases it is advisable to give a general test first and then supplement it, in the case of those pupils who exhibit general shortcomings, with a more detailed test.

Analytical diagnosis. A diagnostic test can do no more than yield detailed information concerning a pupil's achievements. In most cases a study of his scores does not tell us why he failed to achieve or what particular errors he has made. In planning effective remedial instruction it is frequently helpful to secure this information by means of analyzing the pupil's performance. This is going beyond measurement in the sense implied by the ordinary use of educational tests. An analytical diagnosis, however, is frequently profitable, particularly in the case of those pupils who are

found to be lacking in achievement. It may be based upon the performances recorded in response to a standardized educational test, or it may be based upon performances secured in other ways.¹

5. *Useful remedial measures*

Planning remedial instruction. A detailed consideration of remedial instruction is beyond the scope of this volume. When a pupil fails to achieve up to the norm for his mental age he does so for some reason. Barring physical defects and outside influences, the cause is in the instruction he has received. He may need to be given a stronger motive. He may need more instruction, or another kind of instruction. It may be that he has not mastered properly some of the prerequisites for this particular subject. Whatever the causes are, the teacher's problem is to remedy the undesirable condition. Measurement by means of standardized objective tests can only reveal the pupils who are lacking in achievement. They do not reveal the causes. To ascertain the exact cause, and to formulate remedial instruction which will bring the pupil up to the norm for his mental age, will tax a teacher's resourcefulness and his acquaintance with the principles of instruction. No detailed directions can be given. The teacher must rely upon his acquaintance with the process of learning and upon his experience. The tests only reveal the pupils who have need for remedial instruction.

Diagnosis of pupils with reference to their study procedure. The diagnosis of pupils with reference to study procedure is not fundamentally different from diagnosis with

¹ Monroe, W. S. *Measuring the Results of Teaching*, pp. 138-52. Houghton Mifflin Company, 1918.

This gives a summary of several reports of analytical diagnosis in the operations of arithmetic.

reference to achievement. In fact, ability to study effectively is one form of achievement. As we indicated in Chapter III, in the upper grades and in the high school the diagnosis of pupils with respect to their study procedures to a large extent takes the place of diagnosis with respect to achievement. In the diagnosis of study procedure one is concerned with locating the pupils who have not acquired effective methods of study. It is not a question of determining who has achieved and who has not, but who has acquired a procedure which makes achievement possible and who has not acquired it. For example, one might have acquired a good procedure for studying mathematics but at the same time have learned very little mathematics, because he had studied mathematics very little or not at all. In addition to being a prerequisite for achievement, effective methods of study may be considered as objectives. They are among the most important outcomes to be realized by the school.

Relatively few tests are available for securing the information which is necessary in diagnosing pupils with respect to their study habits. Since reading is the basic activity in a number of fields of study, a silent reading test will throw some light upon pupils' study procedures.

Setting immediate educational objectives. General statements of educational objectives are not sufficient to provide pupils with definite goals for their daily work. There is need for immediate and detailed educational objectives that can be understood by the pupils. In so far as the norms derived for our educational tests represent degrees of ability that should be attained, they may be used as objectives. Since they are stated in definite terms, and in the case of many of our tests are given in detail, they naturally become a very satisfactory statement of immediate educational objectives.

For example, if we assume that the norms for the Courtis Standard Research Tests, Series B, represent the degrees of attainment which we should expect of pupils in the operations of arithmetic with integers, these norms may then be used as immediate educational objectives in this field. They can be easily understood by the pupils. They know that their goal is to do so many exercises of a given type in a given time, with a certain per cent of their answers correct. Similar norms can be stated for handwriting, reading, and in fact for all subject-matter fields in which we have objective tests whose norms may be considered to represent satisfactory objectives.

Motivating school work through the use of standardized objective tests. The setting of definite immediate objectives is an effective method of creating a motive. When a group of pupils is given a rate of 200 words per minute as an immediate objective in silent reading they are able to understand what the objective means. They have something very definite to work for. A similar condition exists when they are given an objective in handwriting, such as a rate of 80 letters per minute with a quality of 60 on the Ayres Handwriting Scale. The mere fact that they are given objectives which they can understand and which are definite accounts for the motivation in part. This motivation is, however, increased by a knowledge that instruments are available which can be used to measure their achievement in any stage of their learning. In the case of some subjects, such as handwriting, these instruments may be placed in the hands of the pupils, and they can ascertain from time to time exactly where they stand with reference to the objectives that have been set.

When definite immediate objectives are set by a teacher, and an effort is made to have the pupils attain these objectives, unusual gains in achievement are frequently made.

The following illustration is typical. Just after the mid-year promotion Monroe's Silent Reading Tests were given to the pupils in grades 3-8 inclusive in one building in a large city. Both Forms 1 and 3 were given at this time.

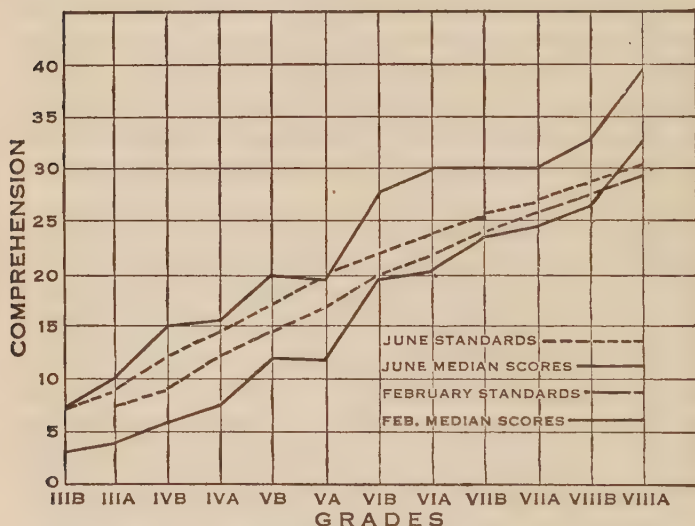


FIG. 6. GAINS IN COMPREHENSION OF SILENT READING DUE TO DEFINITE INSTRUCTION FOR THIS PURPOSE

The median scores showed that in most grades the average attainments of the pupils in silent reading were below the norms for these tests. The teachers of this school adopted the norms of the tests as educational objectives. Without devoting any more time to the teaching of reading, a definite effort nevertheless was made to train the pupils in rate and comprehension of silent reading. Just before the close of the school year, in June, Forms 1, 2, and 3 of Monroe's Standardized Silent Reading Tests were given to these pupils.

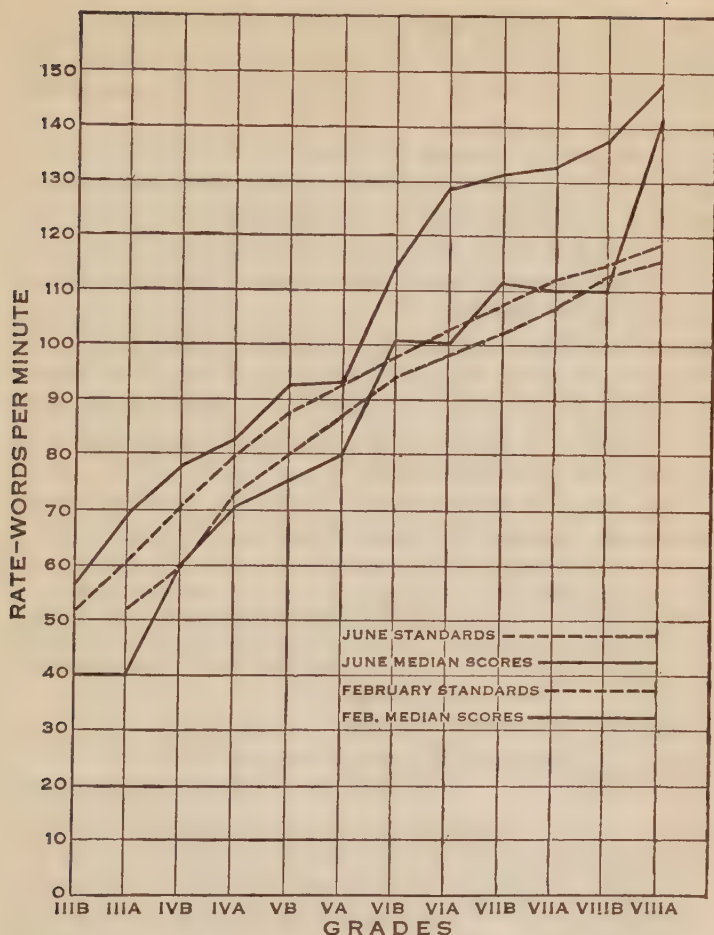


FIG. 7. GAINS IN RATE OF SILENT READING DUE TO DEFINITE INSTRUCTION FOR THIS PURPOSE

In Figures 6 and 7 the average median scores ¹ and the standard scores both for February and June are represented

¹ Although the three forms of Monroe's Standardized Silent Reading

graphically. The average median scores are represented by solid lines in both figures; the lower one being for February and the upper one for June. The broken lines represent norms for February and June.

These two figures show that in each grade relatively large gains were made in both rate and comprehension. These gains in general are considerably greater than the normal gain for the period between February and June. In some of the grades it is four or five times as great. The gains shown were doubtless affected somewhat by the increased acquaintance of the pupils with the tests and with testing conditions. There is no doubt, though, that the adoption of the norms for these tests as educational objectives greatly motivated and directed the school work, so that unusual progress was made by the pupils in the field of silent reading.

Reporting the achievements of pupils to parents. Since scores are more accurate measures of achievement than the

Tests were designed to be equivalent, investigation has shown that they are not equivalent. Pupils make somewhat higher scores upon Forms 2 and 3 than they do upon Form 1. It is estimated that multiplication by the following factors will reduce scores made on these forms to the equivalence of Form 1 scores:

Form 2, Comprehension Test I	.97	Test II	.94
Rate " I	.88	" II	.83
Form 3, Comprehension Test I	1.00	Test II	.88
Rate " I	.80	" II	.93 in grades 6 and 7. .95 in grade 8.

The median scores for Forms 2 and 3 were multiplied by these factors before they were averaged.

In reporting these scores graphically it was necessary to recognize the fact that grades III-B to V-A were given Test I, and Grades VI-B to VIII-A were given Test II. Although it was intended that these two tests should have the same size units and a common zero point, experience has shown that they do not have. It is estimated that in order to make the comprehension scores obtained from Test II comparable with those obtained from Test I, three units must be added to the Test II scores. In the case of rate, it is necessary to add 10 units. This has been done in drawing the graphs so that the lines would show uniform development.

usual school marks, it has been proposed that the scores yielded by the standardized educational tests be entered upon the reports sent to parents. As we have previously pointed out, measures of achievement can be interpreted only by comparison with appropriate norms. Thus, if the scores yielded by standardized educational tests are entered upon the reports sent to parents it will be necessary to enter also appropriate norms. This can be avoided to some extent if the derived measures, described in Chapter VII, are used instead of the point scores. In any case it will be necessary to supply patrons of the school with the information needed for an intelligent interpretation of the scores.

It should be remembered that our standardized objective tests do not yield perfectly accurate measures of achievement. Although they do not appear to involve as large errors of measurement as ordinary school marks, it is important that both children and parents should not be given the impression that the scores are perfectly accurate measures of achievement. If the scores are presented as being ideal measures of achievement both pupils and parents will be inclined to make misleading interpretations. It has been noted in several connections that measures of achievement cannot be completely interpreted without a knowledge of the pupil's general intelligence. There is some doubt whether at the present time it is desirable to report measures of general intelligence to children and their parents. For this reason the use of the results yielded by standardized objective tests should be used with caution in reporting the achievements of pupils to parents.

II. USES OF STANDARDIZED TESTS BY THE SUPERVISOR

Six major supervisory uses for tests. The major activities in which the supervisor may use standardized educational tests are the following:

1. Promotion and classification of pupils.
2. Educational and vocational guidance.
3. Supervision of the instruction including the evaluation of school efficiency.
4. Rating of teachers. (This is partly included in the above activity.)
5. School publicity.
6. Scientific experimentation.

In the first two of these activities the supervisor has to deal with individual pupils. Therefore, he will have need for standardized educational tests that will yield reliable individual measures. In the other four activities mentioned the supervisor is dealing for the most part with groups of pupils. When he wishes to obtain only a general survey or average measure of the school system, or some division of it, he may use a test which does not yield accurate measures of individual pupils. Since the probable variable error of measurement is inversionally proportional to the square root of the number of pupils in the group, the measure of a group of pupils will always be more accurate than the measure of a single pupil. Certain very brief tests have been designed for the use of supervisors. This is true of the Courtis Supervisory Tests in Arithmetic, Monroe's General Survey Scale in Arithmetic, the Indiana Scales of Attainment, No. 2 and No. 1, and the Pintner Survey Test.

1. Promotion and classification of pupils

Grade placement of pupils entering school. When pupils enter school for the first time they are usually placed in the first grade, and all are required to do the same work. Recently it has been suggested that when a school can be organized so as to permit of it, the pupils entering should be grouped according to their capacity to learn. This would mean that some would be placed in I-A, others in II-B, and

possibly some in II-A. In one school system, that has come to the attention of the writer, pupils below a certain chronological age are not admitted to school until they have attained a certain mental age. In placing pupils who are entering school, general intelligence tests are almost indispensable. Since the pupils are just entering the school the supervisor has very little information concerning them. A few minutes devoted to the application of a general intelligence test will enable the supervisor or the teacher to place such children with surprising accuracy.

Placement of transfer pupils. Pupils frequently transfer from one school to another and when the courses of study followed in the two schools are not identical a problem frequently arises in placing such pupils. Standardized educational tests can be used to measure their achievements. A general intelligence test will yield a measure of their capacity to learn. Comparison of the measures with the grade norms will indicate where a pupil should be placed. Of course any additional data, such as his school record, chronological age, health, etc., should be taken into consideration.

Promotion of pupils. In practically all school systems pupils are promoted from one grade to another at specified intervals. The rules governing the promotion of pupils is a matter of school policy. What the passing mark should be; whether the promotion is to be made upon the basis of capacity to learn or upon the basis of past achievements; whether promotion will be on trial or shall be based solely upon previous work; what shall be done with doubtful cases; these are questions to be settled by the supervisor and the instructional staff of the school. No standardized educational tests can answer any of these questions of school policy, but after the general policy has been defined standardized educational tests are valuable instruments in dealing with some

of the problems that arise in connection with the promotion of pupils. They have frequently revealed bright pupils who had not been suspected of this capacity by their teachers. They yield valuable supplementary information in doubtful cases.

TABLE X. PER CENT OF PUPILS FAILING IN ONE LARGE BUILDING OF A CITY SCHOOL SYSTEM

<i>Subject</i>	<i>Per cent of failures</i>	
	<i>6B</i>	<i>6A</i>
Reading.....	20	7
Arithmetic.....	35	36
Spelling.....	10	13
Penmanship.....	42	13
History.....	31	24
Geography.....	28	27
Oral Composition.....	35	22
Written Composition.....	21	29

It is important that there shall be uniformity with respect to the promotion of pupils within a school system. However, it not infrequently happens in a large city school system that some teacher or group of teachers may not conform to the general policy. Standardized objective tests furnish a means for investigating the promotion record of a teacher or of a school. For example, in one school building in a large city school system the per cents of failures were very high in VI-B and VI-A. They are given in Table X. Because such a large proportion of the pupils in these two half grades received a mark of failure the supervisor naturally wondered whether the pupils in this school were so inferior in capacity to learn that an unusually large per cent of them should fail.

In order to answer this question the Illinois Examination, including tests on general intelligence, the operations of arithmetic, and silent reading, was given to the pupils. The scores obtained showed clearly that the pupils in these two

half grades were on the average somewhat superior to pupils in the corresponding grades in other schools. They were shown to be superior not only in general intelligence but also in the operations of arithmetic and silent reading. The median achievement quotients were distinctly above 100, and also above the medians in other schools in this city. In fact these standardized educational tests showed that the pupils in these two half grades were not inferior in either capacity to learn or in their achievements. Consequently, the supervisor was able to assert that the high per cent of failures was not in agreement with the general promotion policy of that school system.

Classification of pupils within a grade. Pupils belonging to the same school grade are known to differ widely with respect to their capacities to learn. Some are dull, others are bright. Even when pupils are placed in the grades that correspond to their mental ages, there will still be differences in I.Q.'s. Those having high I.Q.'s will be able to learn more rapidly than those who have low I.Q.'s. In school buildings where there are two or more teachers giving instruction within the same grade, and to a limited extent in smaller buildings, it is possible to classify the pupils belonging to a school grade for purposes of instruction.

It has been proposed that the efficiency of a school will be increased if the pupils are classified so that those possessing approximately the same degree of brightness will be instructed together. The instructional groups will then be more homogeneous. Approximately the same work may properly be expected from all members of the class. There will be no bright pupils to be bored because of the slow and monotonous pace that is set for the less capable pupils, and there will be no dull pupils who are unable to keep up because they do not have the capacity to learn as rapidly as the average of the class. This proposal is supported by a

great many supervisors and teachers who have put it into practice.

General intelligence tests, supplemented by achievement tests, afford a basis for making such a classification of pupils. The classification should not be made on the basis of general intelligence alone, because it is necessary to consider the attitude of a pupil toward the work of the school as well as a number of other factors which affect his achievement. A pupil's mental age and intelligence quotient are, however, the two most significant items with reference to the classification of pupils for instructional purposes.

Selection of exceptional children. Some children differ so greatly from the average child of the same chronological age that special instruction is required. Exceptionally dull or backward children need to be placed in special classes. Pupils who are exceptionally bright also demand special instruction. In fact the provision of appropriate training for such children is much more profitable to society than the training of dull or backward children. For the selection of both types of exceptional children general intelligence tests are indispensable. Tests for measuring the achievement of pupils, particularly in the case of exceptionally bright children, are also very valuable.

2. Educational and vocational guidance

Use of tests in this work. It is now the consensus of opinion that the school should place the pupil and guide him in the selection of his work so that he will secure the maximum returns for the time he spends in school. The promotion and classification of pupils, which has been discussed above, is one form of educational guidance. However, educational guidance is coming to mean more than this. Terman and others have shown that pupils below a certain mental level have very small chances for success in

certain school subjects. In the elementary school there is little opportunity for the selection of subjects to be studied below the sixth grade. Practically all pupils must devote their time to the study of minimum essentials. Here educational guidance is confined almost exclusively to the promotion and classification of pupils. In the junior high school, and to a greater extent in the senior high school, some election of work is possible.

It is not expected that all pupils will study the same subjects. In making their selections pupils should be advised with reference to the courses which they may undertake with a reasonable chance of success. They should be guided in part by their interests, but it is unwise for pupils to undertake subjects in which the chances for success are very limited. This guidance should be based upon definite information concerning a pupil's capacity to learn, and upon his achievement in the past. Standardized educational tests are valuable instruments for this purpose. The scores obtained should be interpreted with reference to success norms.

3. The supervision of instruction

Objective measuring substituted for subjective estimating. A supervisor must keep in touch with his school system. From time to time it is imperative that he make a survey of it. He needs to compare the efficiency of his system with that of other similar systems. He also needs to be informed of the relative efficiency of the units of his own system. Before standardized educational tests were available, such surveys were made by a personal visitation. Obviously they were subjective. Standardized objective tests afford a means for making an objective survey.

The supervisor diagnoses the school system in much the same way that the teacher diagnoses his class. The general procedure is not materially different. It has been

shown that the most effective method of supervising instruction is to ascertain the divisions of the school system which are below the desired standard of efficiency, and to prescribe definite remedies for undesirable conditions that are found to exist.¹ In this way the efforts of a supervisor will be directed toward the meeting of specific needs. In the absence of a diagnosis of his school system the supervisor will frequently direct his energies in ways that are not profitable. It may be that a reorganization of the system is needed. The course of study may be lacking. The teachers may need some special training. Whatever the needs are the supervisor will be in a better position to meet them when he is acquainted with the conditions that exist.

Evaluating the efficiency of a school system.² Standardized educational tests are sometimes mentioned as instruments which can be used to measure the "efficiency" of a school system, or some division of it. Such statements are misleading. A standardized educational test may be used in arriving at a measure of "efficiency," but one test, or even a group of tests, cannot yield direct measures of it. "Efficiency," as used in education, has a meaning which is not essentially different from that which the term has when applied to an industrial project, or even to a single machine. In these fields "efficiency" is a ratio whose maximum value is 1.00. This ratio is a fraction, whose numerator is the output and whose denominator is the input or investment. In education, the output consists of the changes produced in the pupils; i.e., the controls of conduct that the school engenders. The output for the year is the total of all the

¹ Courtis, S. A. "Measuring the Efficiency of Supervision in Geography"; in *School and Society*, vol. x, pp. 61-70 (July 19, 1919).

² The reader should distinguish between the "efficiency of a school system" discussed here and the "efficiency of instruction" considered on page 172. The term "efficiency" as used here has a more complete meaning.

changes¹ that have been produced in the pupils during the year, due to the influence of the school. The educational investment includes many factors, such as buildings, equipment, textbooks, teachers, supervision, and general administration. It also includes the time during which these investmental factors operate. The ratio is also affected by the quality of the pupil material, organization of school system, methods of instruction, and so forth.

A standardized educational test measures only the achievements of pupils in the field of a school subject. If we assume that the achievements measured by a given test are representative of the total field of a school subject, we may consider that we have an average measure of all of the pupils' achievements in this field. The general intelligence of the pupils must also be measured. Even when we have measured this, we still do not have a measure of the efficiency of the school system, because no account has been taken of the investment in buildings, equipment, textbooks, teachers, supervision, and general administration, and of the time during which these factors operated. Hence, a single standardized educational test, or even a group of standardized educational tests, cannot yield, directly, measures of the efficiency of a school system.

Efficiency measurement calls for intelligence measurement also. The necessity for securing measures of general intelligence of pupils is emphasized in the following illustration. Monroe's Standardized Silent Reading Test, and Monroe's General Survey Scale in Arithmetic, were given throughout two school systems. The two cities concerned are approximately the same size, and do not differ markedly with respect to the character of their population. The median grade achievements of the two cities are shown in Fig. 8. It is clearly evident in this figure that City B is dis-

¹ A more exact statement would be the social worth of these changes.

tinctly superior to City A in both arithmetic and reading, when compared grade for grade. The superiority is more pronounced in the case of reading. If this were the only information at hand for the two cities we would have to conclude that the school system of City B was distinctly superior to that of City A, and that City A was distinctly below the city average.

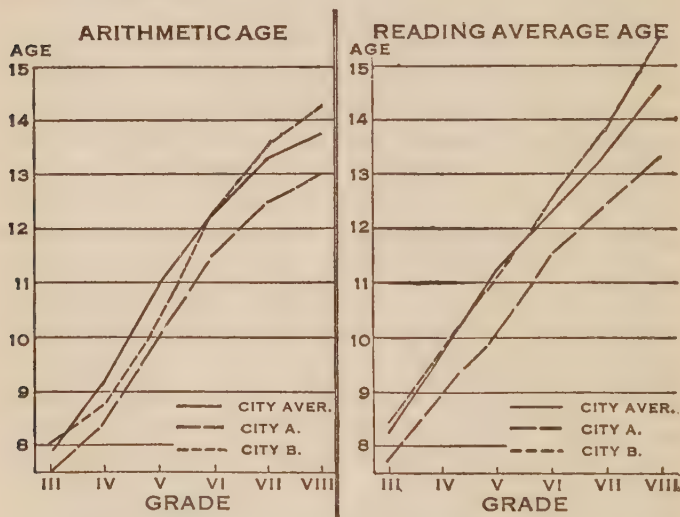


FIG. 8. ACHIEVEMENT AGES FOR ARITHMETIC AND FOR READING (AVERAGE) FOR CITY A AND CITY B

Fortunately, the Illinois General Intelligence Scale was administered to all pupils at the same time that the reading and arithmetic tests were given. The median mental age and the median I.Q. are given for these two cities in Fig. 9. This figure shows that the school system in City B has been organized so that in each grade the median mental age for the pupils is about one year in advance of the median mental age of the pupils of the corresponding grade in City

A. The median I.Q.'s are also distinctly higher. In fact in City B the median mental age of the pupils is shown to be distinctly above the city average. These facts mean that, grade for grade, City B has pupil material that is distinctly superior to that in City A. This is probably not due to any

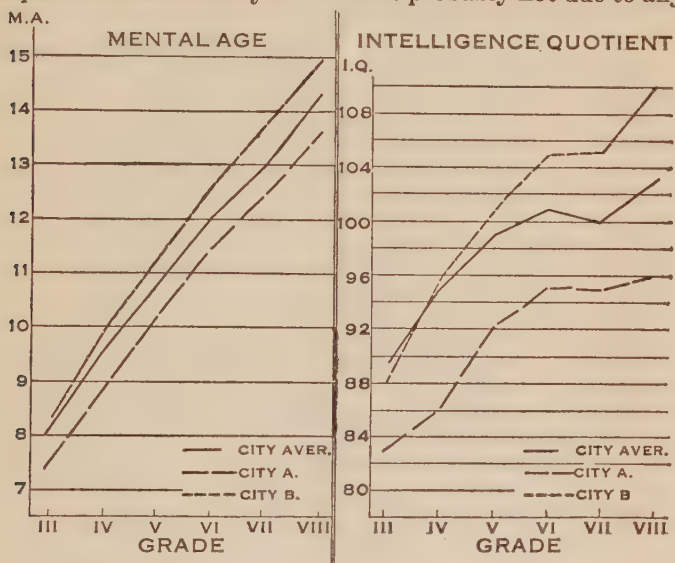


FIG. 9. MEDIAN MENTAL AGE AND MEDIAN I.Q. FOR EACH GRADE FOR CITY A AND CITY B

inherent differences in the quality of the pupils in the two cities, but rather to the differences in the organization of the two school systems. In City B the school system has been organized so that the pupils, particularly in the upper grades, are highly selected. The promotion rate has been relatively low, and pupils who did not get along well in school have dropped out after passing the age limit of compulsory attendance. The opposite policy has predominated in City A. This system has been liberal in regard to promotion, and

there has been an effort to keep pupils in school. This has resulted in less highly selected groups of pupils, particularly in the upper grades.

In Fig. 10 the median achievement quotients are shown for these cities. In the case of arithmetic, the median

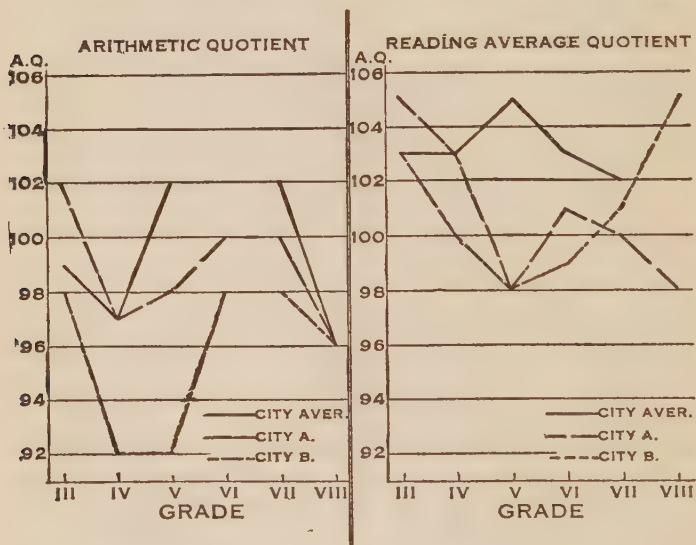


FIG. 10. ACHIEVEMENT QUOTIENTS FOR EACH GRADE FOR CITY A AND CITY B

achievement quotients for City B are distinctly below those for City A, except in the eighth grade. They are also below city average. In the case of silent reading, City B surpasses City A only in the seventh and eighth grades. Therefore, we conclude that, with the exception of the eighth grade and in part the seventh, the school system of City B is less efficient in teaching the operations of arithmetic and silent reading than the school system of City A. This conclusion is the opposite of the one suggested by the absolute

measures of achievement. By having at hand a measure of the mental ages of the pupils, it has thus been possible to avoid an erroneous interpretation of the measures of achievements for these two cities.

The need for measures of the general intelligence of pupils in interpreting measures of their achievements exists also in studying a division of a school system. In fact the need is even greater, because the quality of the pupil material in different sections of the city is likely to exhibit greater differences than we find between cities.

When interpreting the measures yielded by standardized educational tests with reference to the efficiency of the school, it is necessary to bear in mind the various factors which affect the achievements of pupils. General intelligence, as we have just shown, is a very potent factor. Heredity, maturity, and general training affect achievements of pupils as well as the specific training which the school affords. Nationality, race, environmental conditions, and perhaps some other factors must be considered in certain cases. Heredity, maturity, general training, nationality, and race will generally be sufficiently accounted for in the measures of general intelligence, but under certain conditions they may require special consideration. Courtis ¹ has shown that maturity and general training appear to be the most potent factors affecting the development of certain abilities. If this is true, the measurement of such abilities can give no indications of the efficiency of the training which the school provides.

The use of comparative data. In interpreting the median scores for a city comparisons are frequently made with similar scores from other cities. In our discussion of norms, in Chapter VIII, we urged certain cautions in their use. In

¹ Courtis, S. A. *Measurement of Classroom Productions*, pp. 357 ff. Report of Gary School Survey, General Education Board, 1919.

the use of other comparative data for judging the relative efficiency of a school system it is even more important that these cautions be kept in mind. We have just pointed out the necessity of taking into account the general intelligence of the corresponding groups of pupils in the cities concerned. It is also necessary to take into consideration such factors as recency of instruction on the content of the test used, the familiarity of the pupils with the testing procedure, their acquaintance with the particular test used, the teaching time allotted to the subject in which the test is given, and the distribution of the time within this subject.

We have noted in another connection that second-trial scores, when a test is repeated after a time interval of not more than a few days, are likely to be as much as 10 per cent greater than first-trial scores. When a test is repeated after a longer interval no definite statement can be made, because the difference between the scores obtained from the two trials will depend upon the instruction which the pupils have received during this interval. If some of the comparative data represent first-trial scores, while others are from second- or third-trial scores, the data cannot be really considered comparable. Also, corresponding scores from two cities cannot be considered to be comparable unless the corresponding groups of pupils are equivalent in general intelligence, and have received equivalent amounts of instruction. Hence, it is dangerous to compare the median scores obtained for one school with the median scores reported for other schools when nothing is known concerning the factors which materially affect the achievements of pupils.

The measurement of progress. Standardized educational tests have been recommended as instruments which a teacher or supervisor may use to secure a measure of the progress for the year or semester. It has been recommended that a suitable test be given at the beginning of the year or

beginning of the semester, and that this test be repeated at the end of the year, using a duplicate form if possible. The differences between the two sets of median or average scores have been considered to represent the progress made during the interval of time that had elapsed from the first to the second testing. This procedure has been suggested as a method whereby a supervisor might not only keep in touch with the efficiency of his system, but also secure significant measures of the effectiveness of his teachers. A teacher might use it to secure a measure of his own efficiency.

Important modifying factors. The score which the pupil receives on a test depends upon a large number of factors other than his ability. The differences between two sets of median scores are accurate indices of increase in achievement only if all factors other than ability are the same at both testing periods. If the pupils are unacquainted with the test, particularly if they are unacquainted with the testing procedure, this condition cannot exist at the second testing. They will have become acquainted with the test through the giving of it at the first testing period. We urge teachers to use standardized educational tests as a means for diagnosing their pupils. In this connection, it is very properly pointed out that unless something is done to remedy the conditions revealed the giving of the tests cannot have much value. This means that the teachers should take steps to apply remedial instruction where the need for it has been shown. In doing this, the pupils will be made acquainted with the types of exercises which the tests include, unless they did them satisfactorily the first time. Naturally, this will tend to make materially higher scores on the second trial. None of our tests may be considered to be all comprehensive, even though they are representative of the field of a school subject. The teacher who attempts merely to correct the defects revealed at the first testing will have a group of pupils

who will show marked increases. However, these increases are likely not to be representative of the real progress of the pupils in the field of the subject. Another teacher who did not emphasize the remedial instruction to the exclusion of instruction on other topics would show smaller differences between the two sets of scores, but actually might have achieved greater progress.

The recency of instruction also materially affects the scores. If the pupils had received no recent instruction upon the exact content of a test at the time of the first trial, but had received instruction on it just preceding a second trial, large differences would be shown, but they would not be indicative of real progress. In order to offset the effect of recent instruction, it has been proposed that a measure of the progress for a year be secured by comparing the scores made by pupils just after the beginning of one school year, with those made just after the beginning of the next school year. Any increase shown would be a permanent increase, because it would be the residue remaining after the summer vacation. The measurement of progress in this way will probably be more accurate than that secured by repeating the test at the close of the school year.

Although the comparisons of the scores made at the beginning of the school year with those made at the end of a school year, or, better still, with those made at the beginning of the following school year, yield some measure of progress, it is necessary to exercise caution in interpreting these comparisons. It is very easy to do an injustice to a teacher or to an entire school. So many factors affect a pupil's score that these differences may be due entirely to causes other than real growth.

4. The rating of teachers

Standardized objective tests used in rating teachers.

This is essentially a special case of evaluating the efficiency of a unit of a school system, but because of its importance it deserves special treatment. In some school systems increases in salary and promotion depend upon the ratings given to teachers. Since teachers are employed for the purpose of producing certain changes in pupils placed under their instruction, it has been suggested that standardized objective tests may be used to measure these changes, and hence to provide an objective measure of the teacher's efficiency.

According to this proposal, by giving several tests at the beginning of the year a supervisor will be provided with a statement of the present status of the pupils. When the tests are repeated at the close of the school year the differences between the two sets of scores will be an index of teaching efficiency. If an extensive battery of standardized educational tests is used a supervisor will be provided with a measure of the extent to which the teacher has fulfilled his function.

Limitations of the method. This proposal makes a strong logical appeal. Teachers should be judged by the results which they produce. There are, however, some reasons why standardized educational tests should not be given a large place in the rating of teachers. In the first place, we do not have available standardized educational tests for measuring all of the changes which teachers are expected to produce in pupils. We have been most successful in measuring skills, and information in the fields of arithmetic, reading, writing, spelling, etc. These changes in pupils are fundamental, and the teacher who is not effective in producing these changes cannot be counted successful. But there are many other important outcomes. Many of these we are as yet unable to measure. We would be unjust to the teacher, and inconsistent with our educational objectives, if we did not recognize this fact in judging a teacher's efficiency.

The second reason is the errors which may be introduced into our measurements. We have already discussed these under the heading of "measurement of progress." Hence, it is unnecessary to indicate the nature and magnitude of these errors here. It is sufficient to say that the errors may be large enough to distort seriously a rating given to a teacher on the basis of the scores yielded by standardized educational tests. Thus, although standardized educational tests may sometimes be helpful in the rating of teachers, the wise supervisor will use them with caution and with due regard to their limitations.

5. School publicity

Use, and limitations. The measures of achievements yielded by standardized educational tests of pupils lend themselves to publicity. They are definite numerical facts. When they are set beside the corresponding norms, or similar data from other school systems, they will appeal to many patrons of the school. Generally the patrons will be interested, especially, if their school is shown to stand high, or the results of the tests tend to corroborate their opinions.

We have already said enough in other connections concerning the limitations of standardized educational tests to make it clear that a supervisor should be cautious in presenting the scores obtained. In addressing a lay audience it is probably unwise to emphasize the errors involved in our present measurements, but the tests should not be presented as instruments yielding highly accurate measures. To the layman accuracy has a meaning derived from the measurement of physical objects. A test may be highly accurate among tests, but in comparison with instruments for measuring physical objects all of our tests are very inaccurate.

When a supervisor presents the results of a testing program to his constituency he must remember that his audi-

ence is not critical. They are relatively unacquainted with the tests used and their limitations. To them he is vouching for the interpretations as well as the validity of the facts. The way in which the facts and the comparative data which accompanies them are presented influences to a large extent the interpretation. The supervisor should, therefore, be careful to present the facts so that they will be properly interpreted.¹

6. *Scientific experimentation*

Every school is a potential laboratory. Both the teacher and supervisor have an opportunity to determine the value of the methods and devices of teaching, and other items of school procedure. The carrying on of an investigation will tend to arouse the interest of teachers in school, and will also tend to stimulate them to think about their work. Even if little is contributed to the accumulation of human knowledge the work will frequently be very profitable. However, a supervisor should bear in mind that the school exists for the education of children and not to provide a laboratory for scientific experimentation. Occasionally it may be worth while to create artificial conditions for the purpose of studying certain items of school procedure, but frequently it will be unwise to control conditions to the extent that is required for careful scientific experimentation.

Every experiment should have a well-defined purpose. It should be directed toward the solution of some educational problem and not merely for the purpose of satisfying curiosity. Sometimes one or more educational tests are given to a group of pupils without any particular purpose in mind and later an attempt is made to utilize the data collected.

¹ Monroe, Walter S. *Measuring the Results of Teaching*, chap. II. Houghton Mifflin Company, 1918.

This chapter suggests some effective methods of graphical representation.

Such a procedure will seldom lead to results that are significant. In so far as it is possible the experiment should be planned in detail before it is begun. It is necessary to guard against not only the errors which are likely to occur in the scores yielded by the tests used, but also other experimental conditions which may seriously limit the validity of the conclusions.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What is diagnosis? Describe the different kinds.
2. How would you go about choosing a test? What would you need to know in order to make an intelligent selection?
3. Plan a testing program for the elementary schools in a city of 2000. In a city of 5000. In a city of 50,000. Go into detail.
4. Plan a testing program for a high school of 200 pupils.
5. Why is the efficiency of a school not measured by a standardized educational test?
6. Plan an experiment in which educational tests would be used, and indicate the precautions you would take to secure accurate results.
7. Distinguish between "errors of measurement" and "errors of interpretation."
8. To what extent might the results of educational testing be used in a rating scheme for teachers?
9. Can the giving of an occasional test without any particular purpose in mind be justified? If so, on what grounds?
10. Should teachers and supervisors be encouraged to use educational tests before they have become acquainted with their limitations?

CHAPTER XI

THE IMPROVEMENT OF EXAMINATIONS

Examinations not completely replaced by standardized educational tests. In Chapter II examinations set by a teacher or other school official were shown to yield highly inaccurate measures of school achievement, and to be subject to certain other limitations. Beginning with Chapter IV, we have discussed in detail the construction of standardized educational tests in which these defects have been eliminated or reduced to a minimum. These instruments for measuring the achievements of school children have become very widely used, particularly in the elementary school, but examinations set by the teacher remain the most frequently used means of measuring the achievement of pupils. Although properly constructed standardized educational tests are superior in certain respects to examinations set by the teacher, they will probably never entirely replace them as a means of measuring the results of teaching.

At the present time we have satisfactory standardized educational tests for only a few subject-matter fields. In the high school, with two or three exceptions, we have no tests which may be thought of as representative of the entire field of a school subject. Even in the elementary school, except for the tool subjects, we have relatively few satisfactory standardized educational tests which are sufficiently broad in their scope to be used as representative of a school subject. Furthermore, the teacher frequently has need for a measuring instrument adapted to a particular course of study, or to the emphasis which has been given to the subject in the teaching of a particular class. For certain pur-

poses a teacher is justified in departing from the official course of study. When this is done there is need for an instrument to measure the achievements of his pupils. A standardized educational test must of necessity be adapted to generally recognized educational objectives, and to the typical or average course of study. Hence, such a test may not meet the teacher's need. On the other hand, an examination may be constructed by the teacher so as to fit the instruction which the class has been receiving.

It should also be borne in mind that standardized educational tests are not fundamentally different from examinations. In both cases, the pupil is asked to give a performance in response to certain exercises, and then this performance is described by means of a score or "grade." Standardized educational tests are essentially nothing more than refined and improved examinations. Hence, the requirements of test construction should suggest ways in which examinations may be improved. Of course, it will not be possible for the teacher to follow out in detail the procedure of constructing standardized educational tests, but there are certain phases of test construction which may be incorporated in a crude way in the preparation of examinations. Because examinations will doubtless always be widely used for measuring the achievement of pupils, we shall consider certain ways of improving them.

Making examinations more objective. The most serious limitation of the ordinary examination is its subjectivity. In the first place, it is subjective in its administration. The questions are usually written upon the board, and the pupils are required to read them. This practice introduces differences from pupil to pupil. The light may not be good for some pupils; others may have difficulty in reading at a distance. Where it is convenient to do so, the administration of examinations can be made more objective by supplying

each pupil with a typewritten or mimeographed copy of the questions. Occasionally it may be wise to dictate the questions, rather than to write them on the board.

The questions should be stated so that all pupils will interpret them in the same way. If all pupils do not attach the same meaning to a question, the performances which they give will not be truthfully indicative of their abilities. In the construction of standardized educational tests, sometimes exercises that are formulated with care are later found to be ambiguous. A striking illustration occurs in Form 7 of the Army Alpha Intelligence Test. The exercise asks the one taking the test to tell whether "cleave" and "split" mean the same or opposite. Both answers are correct, because "cleave" has two meanings which are exactly opposite. Hence, performances on this exercise are not valid indices of general intelligence. Apparently this ambiguity was not detected, although the test was prepared with care by competent persons. However, in most cases teachers will be able to avoid ambiguous exercises if they subject their questions to a careful scrutiny with reference to this characteristic.

The pupils should be given very definite instructions concerning the method of work to be employed. They should be told whether they are to work rapidly or slowly. In questions in which they are asked to "discuss," the completeness expected in the discussion of the exercise should be indicated. They should be given some indication of the amount of time they will have for the examination, and whether they should answer the questions in order, or not. Other items of procedure in which pupils are inclined to vary should be specified.

Increasing the objectivity of the marking of examination papers by definite rules. The scoring of examination papers has been shown to be highly subjective. There are two ways

in which the subjectivity of the scoring can be greatly decreased. If it is possible to have the examination papers scored independently by two or more teachers, the average of the "grades" assigned to a paper will be more objective than the "grade" assigned by one teacher. The marking of examination papers within a school system can also be made more objective by formulating definite rules. These rules may include some illustrations of the types of answers to be counted as correct, and of those to be counted as wrong. If partial credits are to be given, the rules should include a statement of the types of answers which are to be accepted for the various partial credits. If credit is to be given for correct principle, agreement should be reached concerning the rule to be followed.

F. J. Kelly¹ describes an experiment which is indicative of the increase in the objectivity of marking of examination papers that may be expected from the adoption of uniform rules. Six fifth-grade teachers gave the same examination in arithmetic to their pupils. Each teacher marked the papers for her own pupils, but did not record the marks on the papers. The superintendent then asked a teacher, who was unusually systematic in marking examination papers, to prepare a set of rules to be followed in the marking of these papers. After she had done so, she marked all of the papers in accordance with this plan. Then the teachers who had first marked the papers marked them a second time, following her plan. This provided two marks by the classroom teacher for each paper, the first without following any systematic plan, and the second given in accordance with the rules formulated. Each of these marks was compared with the mark given by the teacher (called the "judge") who marked all of the papers.

¹ Kelly, F. J. *Teachers' Marks*, p. 83. Teachers College Contributions to Education, No. 66.

In Table XI, the six teachers are designated by the letters A, B, C, D, E, and F. The table is read as follows: When

TABLE XI. DISTRIBUTIONS OF DEVIATIONS FROM A "STANDARD" MARK OF TWO SETS OF TEACHERS' MARKS ON FIFTH-GRADE ARITHMETIC PAPERS—FIRST, WITHOUT ANY EFFORT TO UNIFY THE METHODS USED, AND SECOND, BY A COMMON STANDARD (AFTER KELLY)

Range of differences	Without standard							With standard						
	A	B	C	D	E	F	Total	A	B	C	D	E	F	Total
21 or more					2		2							
18 to 20	1				1	1	2							
15						2	2							
14						1	1							
13					1	2	3							
12		1			1		2			1				
11			1		1	2	4							1
10		1				1	1	1						1
9					2	1	4							
8				1	3	1	5							
7	1	1		1	1	1	5				1			1
6		2			1	1	4							
5		1	2	1	1	2	7							
4	2	2	2	1	1	2	10	1					1	2
3		4	2	1	2	2	11			1	1	1		3
2	2	2	1	1	1	1	8	4	1		5	7	1	17
1		5	4	3	2	4	18	2	3	4	5	1	1	16
0	1	4	4	1	1	1	12	22	30	16	16	20	26	139
1	2	5	2	2	2	1	14	5		2	2	1	3	13
2	6	1	3	2	3	1	16	1	1	3				5
3	9		2		2		13		2	2	1		1	6
4	5	1	4	1	5	1	17		2	3	3			8
5	2	3	2	2	1		10		1	1	2			4
6	1	1		3	2		7			1	1			2
7		1	1	6	1		9							
8		2	1	2		1	6							
9	1		1	2			4							
10	1			1		1	3							
11		1	1				2							
12	1			1		1	3							
13		1			1	1	3							
14						1	1							
15			1	1			2				1			1
16 to 20		2					2							
21 or more			1	3	1		5							
Totals	35	41	35	36	39	33	219	35	41	35	36	39	33	219
Medians	+3	0	+1	+6	-1	-4	+1							

no rules were followed teacher A marked one paper 16 to 20 points lower than the "judge," one paper 7 points lower, two papers 4 points lower, two papers 2 points lower, agreed with the "judge" on one paper, etc. The differences between the marks given when the classroom teachers followed no rules and when they followed the rules as formulated are very striking. In the first instance, the marks assigned by the teacher agreed with those assigned by the "judge" in only 5.5 per cent of the cases, while in the second instance they agreed in 63.5 per cent of the cases. This indicates a very marked increase in the objectivity of the marking of the papers.

Increasing the objectivity of examinations by using questions which permit of only one correct answer. The marking of examination papers is subjective largely because the scorer is asked to exercise judgment in determining the credit to be given for the pupil's performance. In spelling, a pupil's performance is either right or wrong, and our practice is to allow no credit for a performance which is not entirely correct. Thus, the marking of an examination paper in spelling is highly objective. A high degree of objectivity may also be obtained in the operations of arithmetic by agreeing to give no credit for examples partly correct. In other subject-matter fields we are accustomed to ask a few questions which call for specific facts, and hence admit of only one correct answer.

It has been claimed that such questions appeal only to the pupil's memory, and that they do not yield an index of his acquaintance with principles and of his ability to organize and apply his knowledge. In order to reach this phase of his education we have asked the pupil to "discuss," "tell why," "compare," etc. When a pupil is asked to formulate an answer consisting of one or more sentences it is difficult or impossible to classify the performance as either right or

wrong. When scorers are asked to exercise judgment in evaluating such performances, wide differences of opinion will be found to exist. In order to overcome this subjectivity of marking, it has recently been proposed that we measure a pupil's acquaintance with principles and ideas by means of certain types of exercises which permit of only one correct answer. These types of exercises have been used in our standardized educational tests, and it is now suggested that they be used by teachers in the examinations which they set. Four types of such exercises will be considered below.

True-false exercises. Instead of asking the pupil to formulate an answer in response to a question, it has been proposed that we ask him to tell whether a given statement is true or false. For example, instead of asking the pupil, "Why did the Puritans come to America in the seventeenth century?" we may ask him to tell whether the following statement is true or false, "The Puritans came to America in the seventeenth century seeking wealth." The pupil may give his answer to this true-false exercise by writing a plus sign if he considers it true, and a minus sign if he considers it false. The mental processes required in answering such exercises do not appear to be the same as those which occur in answering questions of the usual type. However, experimental evidence indicates that there is a very high correlation between the scores which pupils make on a true-false examination, and their acquaintance with ideas and principles as determined by our ordinary examinations.

In constructing true-false exercises, one may prepare a list of statements which cover, in some detail, the portion of the subject on which the pupils are to be examined. After such a list has been prepared, some of the statements can easily be changed to false ones so that the number of true statements will approximate the number of false ones. The

untruth of a statement should not be too obvious, or it will be worthless for testing. An effort should be made to secure statements which will require an acquaintance with the subject in order to determine their truth or falsity. The exercises should be arranged so that there is no regular sequence between true statements and false statements. Since the pupil can give his responses very quickly, the examination should consist of not less than fifty statements. A true-false examination of one hundred statements can be given in the time usually devoted to an ordinary examination.

Procedure in the true-false examination. The examination should be mimeographed or printed so that each pupil will have a copy. He may give his answers in the margin of the sheets, or, if it is desired to use the same set of papers with another group of pupils, he may be given a sheet of paper on which there are numbered blanks. The pupils will then be asked to record in the blanks their answers to the corresponding exercises. A less desirable plan, which may be followed when it is not possible to secure mimeographed copies of the examination, is to read the statements to the pupils and have them record their answers in numbered blanks. The disadvantage of this plan is that the pupils would not have a satisfactory opportunity to study the statements. There is also a chance that the class may give some indication of the answer if a statement should appeal to them as being ridiculous.

The pupils should be given specific directions in regard to answering the exercises when they are uncertain. One writer¹ has suggested that the pupils be instructed to guess, if they were uncertain concerning the truth or falsity of the statement. Another writer² who has used this type of exam-

¹ McCall, W. A. "A New Kind of School Examination"; in *Journal of Educational Research*, vol. I, pp. 33-46 (January, 1920).

² Wood, Ben D. "The Measurement of College Work"; in *Educational Administration and Supervision*, vol. VII, pp. 301-34 (September, 1920).

ination instructed the pupils as follows: "First, go through the list quickly and mark all that you know for certain, then go back and study out the harder ones. Do not guess; the chances are against you on guessing. Don't endanger your score by gambling on those questions about which you know nothing." This is probably the better procedure.

Since only two responses are possible, it is obvious that a pupil may give a correct response as the result of chance. In order to take this possibility into account, a pupil's score on an examination of this type is the number of exercises answered correctly minus the number answered incorrectly. Exercises not attempted are not counted.

"Yes" and "no" exercises. These exercises are just what their name implies. Each exercise is answered correctly by either "yes" or "no." No other answers are appropriate. The administration and scoring are similar to that of the "true-false" exercises. In fact, they may be considered as a special case of the "true-false" exercises.

Recognition exercises. Exercises in which the pupil is asked to choose from a number of proposed answers have also been used to make the scoring objective. This type of question has been called the "recognition exercise." It has been used in standardized silent reading tests, and in a number of our group intelligence tests. It may be illustrated by the following:

"The first president of the United States was: Christopher Columbus, Benjamin Franklin, George Washington, Thomas Jefferson."

A pupil may indicate the correct answer by underlining it, or marking it in some other way. If only one of the answers proposed may be considered correct the marking of such exercises will be highly objective.

Examinations of this type should be mimeographed or printed, and a copy given to each pupil. He should be given

definite instructions concerning the method of work to be followed. It is probably best to instruct him to work through the test rapidly, answering those exercises which he is certain he knows. He should then go back over the list and try the more difficult ones. Not fewer than four proposed answers should be given. When this is done the chances are slight that a pupil will give the correct answer by guessing. The pupils may be instructed to guess if they do not know, since the chance of success by guessing is slight. The pupil's score on an examination of this type may be taken as the number of exercises done correctly.

Completion exercises. In this form of test the pupils may be asked to fill in the words omitted from statements. The following illustrates this type of exercise:

1. Revenue for paying the war debts of the states after the Revolutionary War was provided by a ——— and by ——— due largely to ——— influence.

The slavery question in such states as should be carved from the Louisiana Territory was temporarily settled in ——— by the ———.

The scoring of completion exercises is not as highly objective as the two types mentioned above. Pupils will write a variety of words in the blanks. Different words may have almost the same meaning, and frequently the scorer will be compelled to determine whether the meaning of some word is sufficiently near to that of the correct answer to justify giving the pupil credit for having answered the exercise correctly. In using completion exercises it is necessary to provide each pupil with a mimeographed or printed copy of the examination. The pupil's score is the number of blanks filled in correctly.

Other advantages of the "new examination." In addition to increasing the objectivity of the marking of the examination papers, certain other advantages of the "new

examination" may be noted. There will be a large saving of time for both pupil and teacher. The pupil is called upon to do little or no writing in giving his answers, and he is, therefore, able to respond to a large number of exercises. In scoring there will be little or no occasion to exercise judgment, and the scorer will need only to note the brief responses given by the pupils. In consequence the labor of scoring will be greatly reduced. The saving of time in the giving and scoring will more than offset any additional time that may be expended in the construction of the "new examination."

Another advantage is that the examination can be made more comprehensive. It is traditional for examinations to consist of ten questions. A few are limited to a smaller number, and only occasionally do we find examinations consisting of more than ten questions. The pupils cannot write upon a large number of questions in the time allowed. In consequence the scope of traditional examinations is necessarily narrow. We have stated that true-false examinations should include not fewer than fifty exercises. Examinations consisting of completion exercises or recognition exercises should have a corresponding length. Thus, the "new examination" may be made distinctly more comprehensive than the traditional examination.

Limitations of the "new examination." It does not appear likely that the "new examination," consisting of the types of exercises we have described, will entirely replace the traditional type of examination. The "new examination" cannot be used in mathematics, except to a limited extent. It cannot be used at all in English composition. The following questions taken from Hahn's Scale for Measuring the Ability of Children in History appear to require mental processes distinctly different from those the "new examination" calls for.

' State points of similarity between the position of the United States in 1812 and her position in 1912.

Arrange the following events in order of cause and effect: Force Bill, the Carpetbaggers, Fifteenth Amendment, Negro Rule in Some of the Southern States, Ku Klux Klan.

Name the Presidents of the United States since 1892.

Furthermore, it is likely that pupils would miss valuable experience and training if they were not asked at times to compare, explain, discuss, or define. This is also true of questions in which they are asked to summarize material presented on a topic, or to apply certain principles that have been presented. Hence it is difficult to conceive of the "new examination" being a complete substitute for the traditional examination.

The unequal difficulty of questions not a serious defect. It does not appear that accurate measurements of the abilities of pupils are secured by giving the same credit for answering an easy question as for answering a difficult one. However, investigations of this question, in connection with the scoring of standardized educational tests, have indicated that the errors introduced by this procedure, which appears to be illogical, are not large. After having weighted the exercises of his language and grammar test on the basis of difficulty, Charters dropped the weights because he found that the correlation between the weighted and unweighted scores was slightly over .90.¹

A number of other test-makers have likewise used exercises which were unequal in difficulty without assigning any weighted credits to them. A number of other tests which consist of exercises arranged in ascending order of difficulty have been scored by taking the number of exercises done correctly, which amounts to giving as much credit for doing

¹ Charters, W. W. "Construction of a Language and Grammar Scale"; in *Journal of Educational Research*, vol. 1, pp. 249-58 (April, 1920).

an easy exercise as for doing a more difficult one. A recent study¹ of methods of describing the performances of pupils indicates that the errors introduced by giving the same credit for all exercises are frequently smaller than the variable errors of measurement. Although the unequal difficulty of the exercises undoubtedly introduces an appreciable error in a number of instances, it appears that this source of error, particularly when the examination consists of as many as twenty or more questions, does not cause gross inaccuracies in the examination "grades." Hence, it is probably safe to say that, if extreme differences in difficulty are avoided, and an examination consists of as many as twenty questions, one may safely neglect the errors produced by the unequal difficulties of the questions.

Agreement with minimum essentials. It is obvious that the content of an examination should be in agreement with recognized educational objectives. The questions should be representative of what the pupils should know. Studies of minimum essentials made in connection with the construction of educational tests or otherwise may be utilized by teachers in setting examinations. Such studies as that of Ayres in spelling and Hahn in history afford the teachers valuable assistance in basing an examination on the minimum essentials. Catch questions, or questions reflecting the hobbies of a teacher, have no place in examinations. Neither should questions be asked just because they are difficult, or because the teacher believes the pupils cannot answer them. The only justification for asking questions is that the questions are worth while.

Recognition of significant dimensions. In considering the construction of educational tests it has been pointed out

¹ Monroe, Walter S. "The Description of the Performances of Pupils on Exercises of Varying Difficulty," in *School and Society*, vol. xv, pp. 341-44 (March 25, 1922). See also pp. 128-29.

that the significant dimensions of ability should be measured. In the traditional examination, pupils are generally allowed all the time they desire or are graded only upon the questions that they have time to try, which amounts to the same thing. This means that only the quality of the work is described. In some subject-matter fields, such as operations of arithmetic or handwriting, it is important to have the rate of work described. By setting examinations sufficiently long so that no pupil will finish in the time allowed, and by timing the pupils, it is possible to secure a measure of their rate of work. Measurement in terms of the level of difficulty reached cannot be accomplished except by means of standardized educational tests. Fortunately, the measurement of this dimension of ability is not important in most school subjects.

Norms for ordinary examinations subjective. It is customary to describe the performance of a pupil on an examination in terms of the per cent of the questions which he has answered correctly. School marks are also generally expressed in per cents, or in terms of symbols which are defined in terms of per cents. Passing marks vary; sometimes they are as low as 50 or 60; sometimes one is chosen as high as 75 or 80. The fixing of the passing mark at any per cent such as 60, 70, or 75 is arbitrary. Usually no distinction is made between "scores" in terms of per cent, and "grades" or school marks. If a pupil answers 72 per cent of the questions of an examination correctly he is given a mark or "grade" of 72 per cent. The same passing mark applies to all examinations and to all types of scoring. Thus, the norm or passing mark for a particular examination is largely determined by the difficulty of the examination and the plan of marking which the teacher follows. Hence, it is subjective. A low "grade" may be due either to the pupil's lack of ability or to the high norm set by the

teacher as a result of the difficulty of the examination and the severity of his scoring. Likewise, a high "grade" may be due either to the pupil's exceptional ability or to an easy examination.

"**Grades**" *vs.* **measures of achievement.** This distinction between "grades" and measures of achievement is important. Teachers have not generally been aware of it. One reason for the failure of teachers to recognize this distinction is the fact that both school "grades" and the descriptions of the performances of pupils are generally expressed in terms of per cents. The score which is used to describe a pupil's performance should be distinguished from his "grade." That these are different may be shown by the following illustration.

Suppose that when expressed in terms of per cents the five highest scores for an average class of forty pupils are as follows: 80, 77, 75, 74, 72, and the five lowest are: 49, 47, 46, 44, 41. Suppose, also, that in this school the "grades" are reported in per cents, and that 75 is the passing mark. Since this class is assumed to be typical or average in ability, it is obvious that only a few pupils should receive "grades" below passing. Only three of the scores are up to or above the passing mark. Hence, it is obvious that these scores are different from "grades." They must be translated into school marks, but before this can be done the basis of translation must be determined.

Translating achievement quotients into school marks. In the case of standardized educational tests scores may be translated into school marks by comparing them with approximate norms and with the standard distributions of scores. The general procedure may be illustrated by the plan which has been proposed for interpreting the achievement quotients yielded by the Illinois Examination.¹ When

¹ Monroe, Walter S. *The Illinois Examination*, Bulletin No. 6, Bureau of Educational Research, University of Illinois.

the achievement quotients for the different school grades were assembled together it was found that the distributions were approximately equivalent. Therefore, it was possible to make a single statement with reference to the basis for translating all achievement quotients into school marks. The distribution was divided, so that a certain per cent of the pupils fell into each group. This division might have been made in a number of different ways. The one given below seemed to be the most satisfactory in the case of this particular battery of tests. Instead of descriptive terms, letters or other symbols may be used. For most purposes five groups will be found satisfactory.

<i>Quality of pupil's achievement</i>	<i>Achievement Quotient</i>	<i>Per cent of pupils included</i>
Very superior	{ 165 and above }	1
	{ 135-164 }	6
Superior	117-134	13
Average	83-116	60
Poor	71-82	13
Failure	{ 55-70 }	6
	{ Below 55 }	1

Translating point scores into school marks. The procedure for translating point scores¹ into school marks is similar. It is impossible for a teacher to obtain a standard distribution for an examination in the same way we have obtained standard distributions for our standardized educational tests. The examination which a teacher prepares is given only to one group of pupils, or at most only within a school system. However, crude standards may be determined by making use of the fact that the scores of a large

¹ The term "point scores" is used here to include all descriptions of performances, whether in terms of per cents or in terms of other units. The procedure described applies to point scores yielded by standardized educational tests as well as to examination scores.

number of unselected pupils approximate the normal distribution. It will, of course, frequently happen that a teacher does not have a typical class. It may be that the class is composed of bright children, or that there has been an accumulation of dull children in this particular class. In any class of ordinary size a close agreement with the normal distribution cannot be expected, although when the class consists of twenty or more pupils we generally find that the distribution of measures yielded by a standardized educational test resembles the normal distribution.

A teacher should first determine whether his class is typical or not. The giving of a general intelligence test will be helpful in this connection. A distribution of their I.Q.'s may be considered a very reliable index of the composition of the group. If the median I.Q. of a class is below 100 the teacher may know that he has poor pupil material. If the median I.Q. is above 100 he may know that the class consists of better pupils than the average. If there is a relatively large number of low I.Q.'s it may be expected that there will be an unusually large number of low "grades." Thus, by means of the intelligence quotient and in other ways, the teacher may come to know the general status of his class.

					58	
					56	69
			47	55	69	
	35	44	55	68	75	
27	34	42	52	63	74	
25	32	40	50	60	70	

The point scores of a class should be assembled, as shown above. There are 23 pupils in this class. The median point score is 55. If the class is an average one this median score of 55 should be translated into the median or average grade which the school recognizes. If the grades are reported in

terms of per cents and the passing mark is 75, the average grade will usually be approximately 85. If the class is known to possess superior ability the median score of 55 should be translated into a higher grade. On the other hand, if the class is known to be decidedly below average in ability, 55 should be translated into a lower grade, perhaps as low as 77 or 78. In an extreme case it might even be translated into the passing grade of 75. The translation of the median score into a grade furnishes a basis for translating the other scores.

In general, the lowest scores will be translated into grades below passing. The per cent of pupils who receive grades below passing will naturally vary widely with different classes. There is a somewhat prevalent opinion that the normal probability curve fixes the per cent of pupils who, in the long run, should fail to receive a passing mark. This is a mistaken notion. The normal probability curve tells us nothing concerning the per cent of pupils who should receive any grade until we have determined, from some independent source, the range of ability which we desire each grade to represent. This is a matter of school policy. The school should determine the range of ability to be represented by each mark. The best way to tell this is in terms of the per cents of pupils who, in the long run, will receive each mark. A survey of the present practice would probably reveal that for elementary schools and high schools the per cent of pupils who are failed, varies greatly from system to system. In some cases the per cent is high; in others it is low. It also varies for different school subjects.

Objective norms for examinations. A teacher can introduce a large degree of objectivity in the norms used in interpreting examination scores. The procedure which we have suggested above for translating point scores into school grades, and the definition of the different grades in terms of

the per cent of pupils who are to receive each grade, will tend to make the norms objective. Of course, there will still be a large subjective factor, but gross injustice will be prevented when the teacher happens to set a hard examination, and pupils will be given more appropriate grades when the examination happens to be unusually easy.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. In what ways is an examination different from standardized educational tests?
2. Construct a "true-false examination," and show how you would demonstrate its usefulness.
3. Can we expect that eventually examinations will be completely replaced by standardized educational tests? Why?
4. Are we justified in asking teachers to spend much time in the preparation and use of examinations? Why?
5. Should pupils be excused from final examinations? State your reasons.
6. How would you determine the reliability of an examination? Can you find any account of this having been done?
7. Secure the scores for an examination that has been given to a large number of pupils, and devise a plan for translating these scores into school marks.

CHAPTER XII

STATISTICAL METHODS: THE FREQUENCY DISTRIBUTION AND ITS DESCRIPTION

Purposes of statistical methods. Statistical methods are procedures which are helpful in analyzing and summarizing collections of quantitative facts, and the relationships existing between sets of paired quantitative facts. A small group of facts may sometimes be summarized by crude and unconventional methods, but when one is dealing with several hundred facts, or, as is the case of some investigations, several thousand facts, it is necessary to follow procedures which experience has proved to be effective in dealing with such collections of facts. In addition, the use of "standard" procedures will generally result in an economy of time. It may happen that in some particular investigation a different procedure can be used advantageously, but due caution should always be exercised when departing from tried methods. On the other hand, any procedure which assists in revealing the true significance of the facts should be used unless some recognized method will accomplish the same purpose. There is no virtue in involved and intricate procedures unless they prove more helpful than simpler methods. Other things being equal one should always use the simplest methods that are available.

In order that one's work may be easily understood by others it is advisable to use recognized statistical methods as far as possible. Certain procedures are so generally helpful that for many of them definite rules have been agreed upon and expressed in terms of formulæ. When these procedures

are effective they should be used, and it is best to adhere, even in minor details, to the formulæ which are generally accepted. Failure to do so will not necessarily make our work incorrect. It will, however, frequently result in our being misunderstood, and in discouraging other workers from reading accounts of our work because of the peculiar statistical methods employed.

Knowledge of significance of derived facts necessary. One may use the formula which defines a statistical procedure without understanding the reasons for it, or how it has been derived. Most users of statistical methods are probably unacquainted with the derivation of some of the formulæ which they use. It is, however, imperative that one understand how to interpret the derived facts (median, average, standard deviation, coefficient of correlation, etc.) which these formulæ yield. As has been pointed out above it is the purpose of statistical methods to assist in the analysis and interpretation of quantitative facts. Hence, if the significance of the derived measures is not understood, no progress in the analysis and interpretation of the original facts has been made. A derived fact sustains a definite relationship to the original facts upon which it is based. This relationship is in the nature of a summary or generalization, but there are certain limitations which must be borne in mind.

Sources of error in the derived facts. The original facts are generally not absolutely accurate. This is particularly true when they are scores yielded by educational tests. If the original facts involve errors certain derived facts also will involve errors. In some cases these errors are less for the derived facts, but always due consideration must be given to them.

Many studies are made in an attempt to arrive at the general laws, or relationships, which exist between two quan-

tities, as, for example, the height and weight of children, or achievement in Latin and achievement in English, or general intelligence and achievement in school subjects. Since it is impossible to gather all of the existing facts with reference to such problems, it is necessary that they be studied on the basis of a sample of the total number of facts. Even when a large number of facts have been collected in a random way they may not be entirely representative. It is, therefore, necessary to make allowance for the errors in our derived facts which may be due to the use of a sample.

It is also necessary to bear in mind that, although the facts may be collected according to a procedure which one might expect to give a random sampling, they may not be entirely representative because of the existence of unexpected conditions.¹ For example, the average or median score is used as the grade norm for educational tests. This average or median is based upon a large number of scores of pupils belonging to a given grade. It is assumed that these scores are representative if they have been secured from a large number of cities and the total number of scores amounts to several thousand in each grade. If these scores are obtained from a first trial they will, in general, not be representative of the scores that would be obtained from the second or third trial of the same test. If they are obtained from schools in which semi-annual promotion prevails they may not be representative of scores which would be obtained from schools which have annual promotion. Again, if they have been obtained from schools in which one type of course of study is used, they are not likely to be representative of schools which follow a different course of study. If they are obtained from city schools they will probably not be representative of rural schools.

¹ See "Psychological Examining in the United States Army"; in *Memoirs of the National Academy of Sciences*, vol. xv, pp. 553 ff., for an illustration of the difficulties met in securing a representative sample.

Plan of treatment to be followed. The consideration of the general statistical methods used in the construction, administration, and study of educational tests may be conveniently divided into three sections, (a) grouping facts in frequency distributions, (b) description and use of frequency distributions, (c) relationship between sets of paired facts.

Grouping facts in frequency distributions. In attempting to interpret or summarize a collection of quantitative facts it is helpful to arrange them in a convenient order. If they are arranged in order of increasing magnitude they can be more easily interpreted than when the arrangement is a random one. For example, in Table XII, the scores for rate of silent reading for 81 seventh-grade pupils are given. If these scores were arranged in the order of increasing magnitude their meaning could be more easily grasped.

However, when one is dealing with as large a number of facts as those given in Table XII it is helpful to group simi-

TABLE XII. SILENT READING RATE SCORES OF 81 SEVENTH-GRADE PUPILS AS DETERMINED BY STARCH'S READING TEST NO. 6

108	106	220	94	100	194	246	162	50	150
172	156	94	150	150	100	220	286	168	426
106	158	192	194	460	294	84	220	46	160
108	176	300	260	476	160	194	220	194	220
56	220	340	150	108	308	166	134	248	160
194	134	246	100	194	286	238	246	220	158
268	178	338	194	316	226	220	194	248	226
72	84	220	134	148	152	194	252	220	338
176									

lar facts. In this illustration the facts extend from one score of 46 words per minute up to a score of 476 words per minute. It is difficult to grasp the significance of such an extended array of facts. If the scale is divided into appropriate intervals, and the number of facts falling in each interval is given, we have a summary of the facts which is more convenient to work with and more easily understood

than the original array. In Table XIII the scale from zero to 499 has been divided into intervals of twenty-five units each. These intervals are given in the first column. In the second column the scores falling in each interval are given.

TABLE XIII. SCORES GIVEN IN TABLE XII GROUPED IN INTERVALS TO FORM A FREQUENCY DISTRIBUTION

<i>Scale-intervals</i>	<i>Scores</i>	<i>Frequency</i>
475-499	476	1
450-474	460	1
425-449	426	1
400-424		
375-399		
350-374		
325-349	340, 338, 338	3
300-324	300, 308, 316	3
275-299	286, 294, 286	3
250-274	260, 268, 250	3
225-249	246, 248, 246, 238, 246, 226, 248, 226	8
200-224	220, 220, 220, 220, 220, 220, 220, 220, 220, 220	10
175-199	194, 192, 194, 176, 194, 194, 194, 194, 178, 194, 194, 194, 176	13
150-174	162, 150, 172, 156, 150, 150, 168, 158, 160, 160, 150, 166, 160, 158	15
	152	4
125-149	134, 134, 134, 148	4
100-124	108, 106, 100, 100, 106, 108, 108, 100	8
75- 99	94, 94, 84, 84	4
50- 74	50, 56, 72	3
25- 49	46	1
0- 24		
Total		81

The third column gives the number, or frequency, of scores in each interval. A table consisting of the first and third columns of Table XIII is called a *frequency distribution*. It consists of, first, a statement of the scale intervals, and second, a statement of the number, or frequency, of facts in each interval. Usually the total is entered at the bottom. It is customary to have the zero, or the smallest unit of the scale at the bottom of a frequency distribution table, as in Table XIII. This, however, is a convention and not essential. It is recommended, though, that this plan be followed because it represents the best statistical practice, and fur-

thermore it is in agreement with certain mathematical conventions.

When a quantitative fact is placed in a frequency distribution it loses its identity. All that we know about it then is that it falls somewhere within the interval. Its true value may place it at either extremity, or anywhere between. The frequency distribution tells us nothing concerning the precise value of the fact. In Table XIII there are three quantitative facts in the interval from 300 to 324. These are 300, 308, and 316. In the frequency distribution we know only that three facts are within this interval. We do not know their precise values. The facts which are grouped within an interval may have different values, as in the case of this particular interval, or they may all have the same value, as in the case of the ten facts in the interval from 200 to 224. The frequency distribution is a summary, and details naturally are omitted.

Continuous and discrete series. It is necessary to recognize certain characteristics of the facts to which we apply statistical methods. Most of the facts with which we deal in educational research are quantities in a continuous series.

A *continuous series* is one which is subject to infinite sub-division. A quantity in a continuous series may be expressed to any degree of precision desired. For example, theoretically a pupil's ability to do addition examples might be expressed by a score of 8.79. A pupil's rate of reading might be expressed as 2.385 words per second. When we are dealing with quantities in a continuous series, if the number of facts is sufficiently large, and each is expressed with a sufficient degree of precision, there will be no "gaps" between successive facts when they are arranged in ascending order of magnitude.¹ This condition is seldom, if ever,

¹ Theoretically, irrational numbers are required in order to fill the "gaps" in a series.

realized in actual practice, because the number of facts collected is always finite and they are not expressed with a sufficiently high degree of precision.

A *discrete series* is one in which there are "gaps." For example, if our collection of facts consists of the number of pupils taught by one teacher, we cannot have fractional parts of pupils. If the number of pupils is greater than 25 it must be at least 26. There is a "gap" between 25 pupils and 26 pupils which cannot be closed. If a discrete series is divided into a number of intervals, each including several of the "gaps," we have, so far as statistical procedure is concerned, a series which approximates continuity. Such a series has been called *pseudo-continuous*.

Method of expressing facts in a continuous series. A quantity in a continuous series is generally described in terms of the limits between which it is located. If these limits are close together the fact is expressed with a high degree of precision. If these limits are relatively far apart, the facts are said to be expressed roughly, or with a low degree of precision. In general, more labor is required in making measurements with a high degree of precision, and if the measures are assembled in frequency distributions the precise statements of the facts disappear. For these reasons most educational tests are constructed so that the facts are expressed in terms of the lower limit of the interval in which they fall. Fractional parts of scores are dropped. For example, in the Courtis Standard Research Tests, Series B, a pupil's score for the number of examples attempted is the number which he has completed. If he is working on the tenth example when time is called, and has completed the answer with the exception of the last figure, this example is not counted in his score. His score is 9. In this case 9 means 9 examples entirely completed, plus some fraction of the tenth example. A pupil who has just barely completed

the ninth example receives the same score. In this case a score of 9 means 9 examples completed plus no fraction.

The scores yielded by certain of our measuring instruments are expressed in terms of the point on the scale to which they are closest. For example, a score of 50 on the Ayres Handwriting Scale means that the quality of the pupil's handwriting is nearer 50 than it is to either 40 or 60, or in other words that it is between 45 and 55.

Assumptions concerning frequency distributions. In describing a frequency distribution it is necessary to make certain assumptions concerning the way in which the true values of the facts in each interval are distributed. In the case of a continuous series the usual assumption is that they are uniformly distributed over the interval and that the average value of the interval is that of its mid-point. For example, in Table XIII, the average of the facts in the interval from 200 to 224 is assumed to be 212.5. It is actually 220, or greater, if we consider that the scores are expressed in terms of their lower limits. The average of the quantities in other intervals in this table will differ in most cases from the midpoint of the interval. However, this assumption is the best single assumption which can be made to apply to all frequency distributions, and, in general, calculations based on this assumption give results which closely approximate the results that would be obtained if the calculations were based upon the actual values of quantitative facts. This is particularly true when the number of cases is large.

Questions arising in forming a frequency distribution. In the construction of a frequency distribution it is necessary first to decide upon the intervals to be used. Two questions arise: (1) How many intervals? this includes the question of how large the intervals are to be; and (2) What are to be the limits of the intervals?

No fixed rule can be given concerning the number of in-

tervals which should be used in forming a frequency distribution. The labor of constructing the frequency distribution, and of making statistical computations on it, will be lessened by reducing the number of intervals. However, if the size of the intervals is increased the facts grouped together will differ more widely. Since a fact loses its identity when it is grouped in a frequency distribution, increasing the size of the interval tends to increase the error in derived measures. It is, therefore, necessary to avoid reducing the number of intervals so that significant errors are introduced in the derived measures, and also to avoid a large number of intervals, which would increase the labor of making statistical computations without at the same time materially increasing the accuracy of the derived measures. In general, it is not wise to have the number of intervals exceed 20 nor be less than 10.

In deciding upon the number of intervals it is also wise to bear in mind the convenience of the intervals chosen. The scale of intervals should be such that the labor of assembling the facts in a frequency distribution is reduced to a minimum. So far as possible, one should choose the intervals in agreement with our decimal number system. Intervals of 5, 10, 20, and 25 are much more convenient to use than intervals of 6, 13, 16, 21, etc. In general the intervals should be equal. This, however, is not imperative, but statistical computations are more difficult when intervals are unequal. In a few cases unequal intervals are introduced by the peculiar nature of the scale in terms of which the facts are described. For example, accuracy is usually expressed in terms of a per cent. If we count both zero and 100, which we must do if such measures are included in our collection, we have a scale which cannot be divided into equal intervals. If the scale of intervals, 0-9, 10-19, 80-89, 90-99, is chosen, we still have to provide for 100 per cent. This

is essentially not an interval, but a point on the accuracy scale. In the class record sheet for the Curtis Standard Research Tests in Arithmetic, Series B, the scale of intervals is still further complicated by taking the first interval from zero to 49. The Nassau County Supplement to the Hillegas Scale yields measures in terms of an irregular scale. The values of the compositions of the scale are 1.1, 1.9, 2.8, 3.8, 5.0, etc. A score of 2.8 (or as it is sometimes expressed 3.0) really means that the value of the composition is between 2.35 and 3.3. If the intervals are expressed in terms of their end points they should be as follows for this scale: .55-1.4, 1.5-2.34, 2.35-3.2, 3.3-4.3, 4.4-5.4, 5.5-6.5, etc.

The exact limits of the intervals should be specified or, if they are not expressed, they should be definitely understood. The form used in Table XIII is recommended. The interval is supposed to begin with the first number, and to extend up to and including the second number. If the facts being assembled in the frequency distribution are expressed in fractional form, it is customary to indicate that any fact greater than the second number used in describing the interval but less than the first of the next interval be included in the first interval. This is indicated as follows: 300.0-324.9.

In order to facilitate the publication of frequency distributions the intervals are frequently described in terms of their lower limits. This would mean, in Table XIII, that only the first number would be given, and it would be understood that the interval extends from this lower limit up to but not including the lower limit of the next interval.

Shape of frequency distribution to be expected. Most measures of mental and physical traits form distributions which tend to approximate a standard shape when represented graphically. A typical distribution is one in which there are a few small measures and a few large ones, with

most of the measures near the center of the distribution. The distribution tends to be symmetrical. Such a distribution is called normal. In Fig. 11 each of the three curves represents a frequency distribution of several thousand pupils according to mental age, as measured by the Illinois General Intelligence Scale. These curves closely approxi-

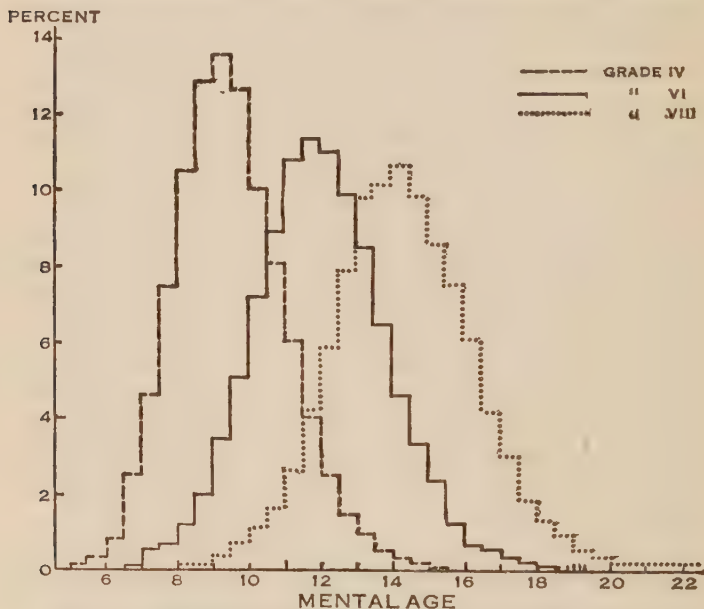


FIG. 11. DISTRIBUTION OF PUPILS ACCORDING TO MENTAL AGE

mate the normal shape, and are typical of distributions formed by accurate measurements of mental and physical traits of unselected groups of pupils. The shape of a frequency distribution depends in part upon the scale of intervals which has been chosen. A different scale may produce a frequency distribution which differs rather strikingly. For example, in Table XIII the scale of intervals is

0-24, 25-49, 50-74, etc. If instead we use an interval of 20, as in Table XIV, a distinctly different frequency of distribution is obtained. Theoretically, this should not happen, but practically it does. It is probably caused by the presence of certain constant errors in some of the rate scores, due to certain peculiarities of the test.

TABLE XIV. FREQUENCY DISTRIBUTION OF FACTS IN TABLE XIII, GROUPED ACCORDING TO A DIFFERENT SCALE OF INTERVALS

<i>Scale-intervals</i>	<i>Scores</i>	<i>Frequency</i>
460-	460, 476	2
440-459		
420-439	426	1
400-419		
380-399		
360-379		
340-359	340	1
320-339	338, 338	2
300-319	300, 308, 316	3
280-299	286, 294, 286	3
260-279	260, 268	2
240-259	246, 248, 246, 246, 248, 250, 7	6
220-239	220, 220, 220, 220, 220, 220, 238, 220, 226, 220, 226, 222, 220	13
200-219		
180-199	194, 192, 194, 194, 194, 194, 194, 194, 194, 194	10
160-179	162, 172, 168, 160, 176, 160, 166, 160, 178, 176	10
140-159	150, 156, 150, 150, 158, 150, 158, 148, 152	
120-139	134, 134, 134	9
100-119	108, 106, 100, 100, 106, 108, 108, 100	8
80-99	94, 94, 84, 84	4
60-79	72	1
40-59	50, 46, 56	3
20-39		
0-19		
Total		81

Description of a frequency distribution. A frequency distribution represents the first step in summarizing a collection of facts. It in turn may be described in terms of (1) its shape, (2) its central tendency, and (3) its variability or spread.

The shape of a frequency distribution. The usual shape of a frequency distribution is one which approaches that of

the normal-frequency curve. A normal-frequency distribution is completely described in terms of its central tendency and its variability. Therefore, we may secure a significant description of a frequency distribution by showing the extent to which it departs from the normal shape. This can be shown graphically by superposing upon the graphical representation of the distribution the normal-frequency curve having the same central tendency and variability.¹

One way in which a frequency distribution may depart from the normal shape is being un-symmetrical. This type of irregularity is called "skewness." Formulas have been proposed for obtaining a numerical statement of the amount of skewness. Thorndike approves the formula,²

$$\text{skewness} = 3 \frac{(\text{mean} - \text{median})}{\sigma}$$

The central tendency of a frequency distribution. The central tendency of a frequency distribution is just what the term implies. It is the "center" of it. Three characteristics of the frequency distribution are used as the central tendency; (1) average,³ (2) median, and (3) mode. Other measures of central tendency could be defined. In fact the harmonic mean is used for certain purposes.

The average. The average is a familiar term. It has long been used in connection with school marks, and even pupils in the elementary schools understand how to compute their "average grade." This is done by adding together the

¹ See H. O. Rugg, *Statistical Methods Applied to Education*, pp. 210-13, for the method of doing this.

² See G. Udney Yule, *An Introduction to the Theory of Statistics*, p. 150, for formula in terms of the quartile deviation.

³ Average is used here instead of "arithmetical mean," which is preferred by some authorities. One reason given for the use of arithmetical mean is that the word "average" is used with a general meaning to include all central tendencies, instead of just one. However, average is properly used in this more restricted sense.

different grades received, and dividing by the number of such grades. The quotient is the average. This method is convenient to use when the number of facts to be averaged is small, but it is a cumbersome procedure when the number is large. It is also not a convenient method to use in connection with a frequency distribution. The method described below is one to be used with frequency distributions. It is sometimes spoken of as the "short" method.

Calculation by the "short method." The procedure for calculating the average is illustrated in Table XV. The frequency distributions is taken from Table XIII. The first step in applying this method is to assume an average. This may be chosen at any point, but the arithmetical calculation will be reduced to a minimum when the assumed average is chosen at the mid-point of the interval in which the true average falls. If it is chosen elsewhere the same results will be obtained, but the labor of calculation will be slightly increased. In applying this method it is necessary to bear in mind the assumption that the average of the measures within a scale-interval is considered to be at its mid-point. In Table XV the average has been assumed to fall at the mid-point of the interval from 200 to 224, or at 212.5. The eight scores in the interval 225 to 249 are 25 units, or one scale interval above this assumed average. In order to reduce the calculations to a minimum the deviation is called one scale-interval. In the same way the deviations of the other scores from the assumed average are expressed in terms of intervals. A negative deviation means that the scores fall in an interval below the assumed average. The deviations are given in the third column of the table. In the fourth column of the table the products of the deviations and frequencies are recorded. The sum of the positive products is 80, and the sum of the negative ones is -132. The difference is -52. This is divided by the total of the

310 THEORY OF EDUCATIONAL MEASUREMENTS

frequencies, which gives a quotient of $-.642$. This means that the assumed average is too large by $.642$ of an interval. Since the interval is 25 units, this quotient is multiplied by 25 to find the correction in terms of units. This correction subtracted from the assumed average gives the true average of 196.45.

TABLE XV. ILLUSTRATING THE CALCULATION OF THE AVERAGE BY THE SHORT METHOD

<i>Scale-intervals</i>	<i>Frequency (f)</i>	<i>Deviation in intervals (d)</i>	<i>Frequency × Deviation (fd)</i>
475	1	11	11
450	1	10	10
425	1	9	9
400		8	
375		7	
350		6	
325	3	5	15
300	3	4	12
275	3	3	9
250	3	2	6
225	8	1	8
200	10	0	80
175	13	-1	-13
150	15	-2	-30
125	4	-3	-12
100	8	-4	-32
75	4	-5	-20
50	3	-6	-18
25	1	-7	-7
0			-132
Total	81		

Assumed average . . . 212.5
 Correction -16.05

 True average 196.45

The average calculated according to the above method will not be exactly equivalent to the average obtained by dividing the sum of the scores by their number. The sum of the scores from which Table XV was derived (see Table XII) is 15,880. This sum divided by 81 gives a quotient of 196.04. The difference is due to the fact that when the scores are assembled in a frequency distribution they are assumed to have the value of the mid-point of the interval in which they are located. By referring to Table XIII it will be seen that this assumed value is not in agreement with the actual value, but since the scores must lose their identity in the frequency distribution it is necessary to make some assumption concerning their magnitude, and this is the best single assumption which can be made.

It has already been pointed out that most scores are expressed in terms of a lower limit. Fractions are generally dropped. For example, if a score of 9 yielded by Courtis Standard Research Tests, Series B, were expressed with a high degree of precision it would appear as 9.43, or possibly 9.429. When the scale-intervals are only one unit, assuming that the scores within any interval are concentrated at its mid-point tends to correct for dropping fractions in the scores. In the above illustration we should expect the average calculated from the distribution to be slightly larger than the one calculated from the scores. It is actually 0.41 larger ($196.45 - 196.04 = 0.41$). In some cases it will be smaller, due to the fact that the scores are not uniformly distributed.

Calculation of average when intervals are irregular. If the intervals of the frequency distribution are irregular in magnitude the calculation of the average is more complicated, although the above method can be applied. It is necessary to express the exact deviation of each interval from the assumed average. Table XVI illustrates the cal-

culatation of the average for one case of irregular intervals. The intervals in this table are equal, except the one for 100. This is really not an interval. It has no width. The average is assumed at the middle of the 50 to 59 interval, or at 55. The distance of 100 from this assumed average is 45 units, or 4.5 intervals. When this is used as the deviation the calculation of the average offers no difficulty.

TABLE XVI. SHOWING CALCULATION OF AVERAGE FOR ONE CASE OF IRREGULAR INTERVALS

<i>Scale-intervals</i>	<i>Frequency (f)</i>	<i>Deviation in intervals (d)</i>	<i>Frequency × Deviation (fd)</i>
100	3	4.5	13.5
90-99	0	4	
80-89	1	3	3
70-79	2	2	4
60-69	5	1	5
50-59	5	0	25.5
40-49	4	1	- 4
30-39	3	2	- 6
20-29	2	3	- 6
10-19	1	4	- 4
0- 9			-20
Total	26		

Assumed average.....55.0

Correction..... 2.1

True average.....57.1

If the scale of intervals has been chosen so that it is incorrect to assume that the measures are concentrated at the mid-point of the interval, due allowance for this must be made in choosing the assumed average. For example, if

the scores obtained from the use of the Ayres Handwriting Scale are arranged in a frequency distribution whose intervals are indicated as 40, 50, 60, etc., the assumed average must be chosen as 40, 50, or 60, and not as 45, 55, or 65, because as this scale is ordinarily used 40 means from 35 to 44, 50 means from 45 to 54, etc.

The median. The median is generally thought of as the middle score, or, more specifically, as the score which has as many scores above it as it has below it. To guide one in the calculation of a median it is necessary to have a more precise statement. A recent writer¹ has given the following definition, "When measures are arranged in order of size the median is the middle measure or (lacking a middle measure) midway between the two middlemost measures." In applying this general definition to some groups of scores which arise in the use of educational tests it is necessary to supplement it by certain arbitrary rules. The median may be calculated from the measures when they are simply arranged in order of size. It may also be calculated from the frequency distribution.

The median of discrete measures. When discrete measures are arranged in order of size it is only necessary to locate the middle measure, if the number of measures is odd. This middle measure is the median. If the number of measures is even there is no middle measure. Then, according to the above rule, the median is midway between the two middlemost measures. This is their average. It is, of course, necessary to correct this average for absurdity when the scores are discrete. For example, it is absurd to give a median in a discrete series as 7.5 children.² The median is either 7.0 or 8.0.

¹ McCall, W. A. "How to Compute the Median"; in *Teachers College Record*, vol. xxi, p. 133 (March, 1920).

² Although it is logically absurd to use such fractions as 7.5 children, the

When discrete measures are assembled in a frequency distribution whose intervals are one unit, all of the measures are concentrated at the lower limit of the interval. It is, therefore, only necessary to locate the interval in which the middle measure falls, if there be one. In case there are two middlemost measures falling in adjacent intervals, the lower limit of either interval may be used as the median. In case the two middlemost measures fall within the same interval the lower limit of this interval is the median.

The median of continuous measures. Calculation of the median of continuous measures is less simple. If the measures are arranged in order of size, it is necessary to estimate the most probable value of the middle measure, or, in case there is no middle measure, of the two middlemost measures. To do this it is necessary to remember that the measures are expressed in terms of lower limits.

For example, the following scores were derived from the Courtis Standard Research Tests, Series B:—2, 3, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9. There are seventeen scores in this group. The middle measure is the ninth one. This is the third 6. The most probable precise values of the four scores which are given as 6 are 6.125, 6.375, 6.625, and 6.875.¹ Thus, the value of the middle measure of this distribution is 6.625. If another measure is added to the distribution to make eighteen there is no middle measure, and we must use the two middlemost measures. Suppose that a score of 10 is added. The two middlemost measures are then the third and fourth 6's. The median is the average of these,

$$\frac{6.625 + 6.875}{2} = 6.75$$

practice may be justified for some uses of derived measures. Frequently the use of fractions will assist in making comparisons more precise. However, the logical absurdity should not be forgotten.

¹ These values are "most probable" according to the assumption of uniform distribution within the interval.

When the two middlemost measures fall in adjacent intervals the median is the junction point between them. This rule is inconsistent with taking the average of the values of the two middlemost measures when the frequencies in the two intervals are not equal. However, the inconsistency is more than compensated for by the ease of calculation which it makes possible.

Calculation of the median from a frequency distribution. The calculation of the median from a frequency distribution is illustrated in Table XVII. The total of the fre-

TABLE XVII. ILLUSTRATING THE CALCULATION OF THE MEDIAN OF THE FREQUENCY DISTRIBUTION. THE SCORES WERE DERIVED FROM COURTIS STANDARD RESEARCH TESTS, SERIES B

<i>Number of examples</i>	<i>Number of pupils</i>
16.....	1
15.....	2
14.....	4
13.....	5
12.....	5
11.....	7
10.....	4
9.....	4
8.....	3
7.....	2
6.....	1
5.....	—
4.....	1
3.....	1
2.....	—
1.....	—
0.....	—
Total.....	40
Crude Median.....	11.00
Correction.....	.57
True Median.....	11.57

quencies is 40. To calculate the median this total is divided by 2, which gives 20, the half sum. We then sum the frequencies, beginning at the bottom, until the partial sum is such that the addition of the next frequency will give a sum greater than 20. In this illustration a partial sum of 16 is reached by the addition of the frequency in the ten interval. The addition of 7 in the next interval will give 23, which is greater than the half sum 20. The addition of the frequency of the ten interval means that we have entirely covered this interval. Therefore, the beginning of the next interval is taken as the crude median. This is 11.0. In order to secure a partial sum of 20 it is necessary to use four of the seven pupils whose scores fall in the eleven interval. In other words, we need to use four sevenths of the cases in this interval. Since they are assumed to be uniformly distributed over the interval, this is equivalent to saying that the median is reached when we progress to a point four sevenths of the width of the interval above 11.0. Since the width of the interval is 1.0, the true median becomes $11\frac{4}{7}$ or 11.57.

We may also calculate the median by beginning at the top of the distribution and counting downward. When this is done the addition of the frequency in the 12 interval gives a partial sum of 17. Since we are counting downward we have progressed to the lower limit of the 12 interval. Hence, the crude median is 12.0. To arrive at the true median we need three of the seven scores which fall in the eleven interval. In other words, we need to progress to a point three sevenths of the width of the interval below 12.0. Hence, the true median is 12.0 minus $\frac{3}{7}$, or 11.57. This is the same result as was obtained by counting upward.

In the above illustration the total of the frequencies was an even number. When it is an odd number the procedure is exactly the same. Table XVIII illustrates the calculation

of the median in such a case. The half sum is 19.5. The maximum partial sum counting up is 14, which carries us to the beginning of the ten interval. In order to secure a partial sum equal to the half sum 5.5 scores within the ten interval are needed. Dividing 5.5 by 9, the frequency in this interval, a quotient of .6 is obtained. Since the width of the interval is 1.0, this is the correction to be added to the crude median. Hence, the true median is 10.6.

TABLE XVIII. ILLUSTRATING THE CALCULATION OF THE MEDIAN WHEN THE TOTAL OF THE FREQUENCIES IS AN ODD NUMBER. SCORES DERIVED FROM THE COURTIS STANDARD RESEARCH TESTS, SERIES B

<i>Number of examples</i>	<i>Number of pupils</i>
14.....	1
13.....	3
12.....	5
11.....	8
10.....	9
9.....	6
8.....	4
7.....	2
6.....	1
5.....	—
4.....	1
3.....	—
2.....	—
1.....	—
0.....	—
Total.....	39
Crude Median.....	10.0
Correction.....	.6
True Median.....	10.6

In both of the above illustrations the width of the interval has been 1.0. When the width of the interval is greater than 1.0 it is necessary to multiply the quotient obtained by

the width of the interval in order to find the correction. For example, in Table XIII, the width of the interval is 25. The half sum for this table is 40.5. The maximum partial sum counting up is 35. To obtain the half sum it is necessary to use 5.5 of the scores falling in the 175 interval. The quotient, dividing 5.5 by 13, is .4. This quotient must be multiplied by 25, the width of the interval which gives 10, the correction. Hence, the median is 175 plus 10, or 185. If one counts downward in this distribution the maximum partial sum is 33 and the required addition is 7.5. The quotient is .6 which makes the correction of 15. The median, therefore, is 200 minus 15, or 185.

Calculation in special cases. Several special cases arise which create difficulties. When the maximum partial sum obtained in counting upward is equal to the half sum, the median is the upper limit of the interval whose frequency was last added, or, what is the same thing, the lower limit of the next interval. For example, in Table XIX a partial sum of 15 is reached by the addition of the frequency in the nine interval. This is also the half sum. Hence, the median is 10.0. If we count downward we use the lower limit of the last interval whose frequency is added.

It may occasionally happen that the two middlemost measures fall on opposite sides of a gap in a distribution. Table XX illustrates this condition. The total of the frequencies is 90. A half sum of 45 is obtained by counting upward to the 180 interval. It is also obtained by counting downward to the 180 interval. In this case the median is the average of the upper and lower limits of this interval, 190 and 180, or 185.¹

The presence of irregular intervals in a distribution do not affect the calculation of the median unless it happens to fall in an irregular interval. Even in such cases it is only

¹ This statement is not precisely accurate. See page 314.

necessary to remember the width of this interval. A peculiar condition which frequently arises in dealing with the scores as tabulated on the class record sheet in the Courtis Standard Research Tests, Series B, is shown in Table XXI. In this case the first frequency is greater than the half sum 10. Hence the crude median is zero, and the correction is ten elevenths of the width of this interval, which is 50. This gives a correction of 45.5, which is the true median.

TABLE XIX. ILLUSTRATING CALCULATION OF MEDIAN WHEN IT FALLS AT THE JUNCTION POINT OF TWO INTERVALS

<i>Number of examples</i>	<i>Number of pupils</i>
14.....	—
13.....	1
12.....	3
11.....	5
10.....	6
9.....	4
8.....	5
7.....	3
6.....	2
5.....	—
4.....	—
3.....	—
2.....	—
1.....	—
0.....	—
Total.....	30

Calculation of percentile points in a distribution. In calculating a median we are simply finding a point in a distribution such that 50 per cent of the measures are on either side. It is possible to calculate the point which represents any division of the frequency distribution. For example, we may calculate the point which has 25 per cent of the measures below it, and 75 per cent above it. Such a point is

TABLE XX. ILLUSTRATING THE CALCULATION OF THE MEDIAN WHEN IT FALLS IN A GAP. RATE SCORES ON STARCH'S SILENT READING TEST NO. 7, GRADE VII

<i>Words per minute</i>	<i>Number of pupils</i>
390.....	1
380.....	—
370.....	—
360.....	—
350.....	—
340.....	—
330.....	—
320.....	1
310.....	—
300.....	—
290.....	1
280.....	1
270.....	3
260.....	—
250.....	1
240.....	—
230.....	2
220.....	3
210.....	6
200.....	12
190.....	14
180.....	—
170.....	6
160.....	5
150.....	3
140.....	2
130.....	4
120.....	14
110.....	1
100.....	2
90.....	6
80.....	2
Total.....	90

called the 25-percentile. The calculation follows exactly the same procedure as that for the median, with the excep-

tion of dividing the total by 2 to find the half sum. To find the 25-percentile we divided by 4 to find one fourth of the sum. The 75-percentile is the point which has 75 per cent of

TABLE XXI. COURTIS ARITHMETIC TESTS, SERIES B,
DIVISION. PER CENT OF ACCURACY

<i>Per cent of attempts correct</i>	<i>Number of pupils</i>
100	3
90-99.....	—
80-89.....	3
70-79.....	1
60-69.....	—
50-59.....	2
0-49.....	11
Total	20

the measures below it, and 25 per cent above it. It is calculated by taking three fourths, or 75 per cent of the total, in the place of the half sum.

Similarly, we may calculate any percentile point, 5-percentile, 10-percentile, 17-percentile, etc. In all cases the procedure is identical with that for the calculation of a median, except instead of finding the half sum, which is really the 50-percentile sum, we find the per cent of the total of the frequencies corresponding to the percentile point desired.

Mode. The mode of a frequency distribution is the value of the interval which contains the largest frequency. No calculations are required in the determination of a mode. It is only necessary to examine the frequency distribution and locate the interval in which the maximum frequency occurs. Some statisticians have attempted to calculate a theoretical mode, which is defined as the mode in the corresponding frequency distribution of an infinite number of cases and infinitely small intervals.¹

¹ See Rugg, H. O. *Statistical Methods Applied to Education*, pp. 100-03. Also, Whipple, G. M. *Manual of Mental and Physical Tests*, pp. 19-20.

Measurement of variability of a frequency distribution. The variability of a frequency distribution refers to the deviation, or spread of the measures, about the central tendency. Some frequency distributions are spread out very widely about the central tendency, while others are closely grouped about it. The difference between the central tendency (the median or average) and a measure is its deviation. If the measure is less than the central tendency, the deviation is negative; if it is greater, it is positive.

The amount of variability of a frequency distribution is measured in several ways. One of the simplest measures is the *average deviation*. It is simply the average of all the deviations. This average is taken without regard to sign. The method of its calculation is obvious, when the measures are not grouped in the form of a frequency distribution. It is simply required to find the sum of the deviations, and divide this by the number of measures.

Calculating average deviation of a frequency distribution. When the measures are arranged in a frequency distribution the calculation is more difficult. Table XXII illustrates the calculation of the average deviation of a frequency distribution. The median of this distribution is 11.57. In calculating the average deviation the arithmetical work will be greatly lessened if we use, instead of the true median, an assumed median at 11.5, the mid-point of the median interval. After we have made our calculations it is relatively easy to correct for the difference between this assumed median and the true median. In the third column of the table the deviation of each interval from the assumed median is given. The fourth column contains the products of the deviations and frequencies. The total to these without regard to sign is 87. This divided by 40, the total of the frequencies, is 2.175.

This, however, is not the true average deviation because

TABLE XXII. ILLUSTRATING THE CALCULATION OF THE AVERAGE DEVIATION OF A FREQUENCY DISTRIBUTION

<i>Number of examples</i>	<i>No. of pupils (f)</i>	<i>Deviation (d)</i>	<i>fd</i>
16	1	5	5
15	2	4	6
14	4	3	12
13	5	2	10
12	5	1	5
11	7	0	—
10	4	-1	-4
9	4	-2	-8
8	3	-3	-9
7	2	-4	-8
6	1	-5	-5
5	—	-6	—
4	1	-7	-7
3	1	-8	-8
2	—	—	—
1	—	—	—
0	—	—	—
Total	40		87

Median = 11.57

it is based upon the assumed median rather than the true median. To correct for this proceed as follows: — The deviation of the 12 interval was taken as 1. Its true deviation is .93. ($12.50 - 11.57 = .93$). Hence, we have used the deviation for this interval which is .07 too large. The same thing has been done for all of the intervals above 12. For the 10 interval the deviation from the assumed median was taken as -1. From the true median it is -1.07. In this case the difference between the assumed median and the mid-point of this interval gave a deviation which was .07 too small. The deviations used for all intervals below the tenth are likewise .07 too small.

This disposes of all the intervals except that in which the median falls. The only difficulty which arises in making the correction is in regard to this interval. It is to be disposed of by the following rule. If the true median is above the mid-point of the interval it is to be counted with the intervals below the median. If the true median falls below the mid-point of this interval it is to be counted with the intervals above. Since the deviations for all of the intervals above the median are in error by the same amount, we may find the total of their frequencies by disposing of the median interval by the rule just given. In this table there are seventeen frequencies which were given deviations .07 too large. There were twenty-three frequencies which were given deviations .07 too small. Stated algebraically, the error in the total deviation is equal to $17(-.07) + 23(.07)$. This is equal to .42. This, added to 87, makes the true total of the products (fd) 87.42. The true average deviation is to be obtained by dividing this total by the total of the frequencies. This quotient is 2.18.

The above procedure may be summarized by the following formula:

$$A.D. = \frac{\Sigma fd + c(N_b - N_a)}{N}$$

In this formula Σfd means the summation of the products of the frequencies or deviations without regard to sign. In Table XXII it is the sum of the numbers in the column headed fd . c is equal to the true median minus the assumed median. N_b is the number of measures below the true median. It is assumed that the measures are concentrated at the mid-point of the interval. N_a is the number of measures above the median. N is the total of the frequencies. When the width of the interval is greater than one it is necessary to multiply this quotient by the width of the interval.

This formula applies only when the correction is equal to or less than one interval.

The calculation of the average deviation has been explained with reference to the median. It may also be calculated with reference to the average. The procedure is identical with that described.

Standard deviation. The variability of a frequency distribution may also be measured in terms of the standard deviation. This statistical term is generally referred to as "sigma" (σ). It is defined by the formula

$$\sigma = \sqrt{\frac{\sum fd^2}{N}}$$

In this formula d represents the deviation of any interval with respect to the true average, f stands for the frequency of measures within the interval, and N is the total of the frequencies of the distribution. When the width of the interval is not equal to one it is necessary to multiply the square root by the width in order to obtain σ in terms of the unit of the measures. The standard deviation is used only with reference to the average.

Table XXIII illustrates the calculation of the standard deviation from a frequency distribution. The procedure, up to a certain point, is identical to that for calculating the average.¹ In the last column of the table the products of the frequency and the square of the deviation of each interval are entered. The total of these products form the numerator in the formula for the standard deviation.

In this table the deviations have been expressed from the assumed average 212.5. The true average is 195.45. It is, therefore, necessary to correct for the error introduced by this assumption. This error is -.642 of an interval. The

¹ See page 309.

TABLE XXIII. ILLUSTRATING THE CALCULATION OF THE STANDARD DEVIATION OF A FREQUENCY DISTRIBUTION. THE MEASURES ARE RATES OF READING, AS MEASURED BY THE STARCH SILENT READING TEST

<i>Scale-intervals</i>	(<i>f</i>)	<i>d</i>	<i>fd</i>	<i>fd</i> ²
475	1	11	11	121
450	1	10	10	100
425	1	9	9	81
400		8		
375		7		
350		6		
325	3	5	15	75
300	3	4	12	48
275	3	3	9	27
250	3	2	6	12
225	3	1	3	3
200	10	0	0	0
175	13	-1	-13	13
150	15	-2	-30	60
125	4	-3	-12	36
100	8	-4	-32	128
75	4	-5	-20	100
50	3	-6	-18	108
25	1	-7	-7	49
0				
Total	81		-132	966

$$\sigma = \sqrt{\frac{966}{81} - .642^2}$$

$$= \sqrt{11.5138}$$

$$= 3.40$$

$$\sigma \text{ (in terms of units)} = 85.0$$

procedure for making the correction is indicated by the following formula:

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - c^2}$$

In this formula *d* is the deviation of the mid-point of an interval from the assumed average. Since *d* and *c* are

expressed in terms of intervals, it is necessary to multiply the value of sigma obtained from this formula by the width of this interval, if it is desired to have the standard deviation expressed in terms of the unit.

Other measures of the variability of a frequency distribution. One measure of the variability of a frequency distribution is half of the difference between the 75-percentile point and the 25-percentile point. This is called the quartile range. The procedure of calculation is obvious. It is simply necessary to find the 75-percentile point and the 25-percentile point, subtract the one from the other, and divide the difference by 2.

The median deviation, or probable error (*P.E.*). In case the distribution is approximately normal, the quartile range can be interpreted as a measure of deviation from the median. It is, thus, essentially the median deviation. This measure of deviation has been called the probable error (*P.E.*). When used in this connection the term is somewhat misleading, and some authorities advise against its use. The median deviation (or *P.E.*) can be calculated directly in case the distribution is approximately normal, but in general the simplest procedure is to derive it from the standard deviation. The procedure is indicated by the following relationship between these two measures of deviation:

$$P.E. = .6745 \sigma$$

This relationship assumes a normal distribution or one which deviates from it very slightly.

Relationship between measures of variability. When a distribution is normal the quartile range is the same as the median deviation, which is the probable error (*P.E.*). The standard deviation, average deviation, and *P.E.* are definitely related to each other. The various relationships are given by the following equations:

$$\sigma = 1.2533 \text{ A.D.}$$

$$\sigma = 1.4826 \text{ P.E.}$$

$$\text{A.D.} = .7979 \sigma$$

$$\text{A.D.} = 1.1843 \text{ P.E.}$$

$$\text{P.E.} = .6745 \sigma$$

$$\text{P.E.} = .8453 \text{ A.D.}$$

These relationships hold only when the distribution is normal, or approximately so. They may be used to obtain any two of the three measures of variability when the third one is known.

Interpretation of measures of variability. The median deviation, or probable error, permits of the most simple interpretation. If we think of a distance equal to the probable error being measured on either side of the average, and the points $Av. - P.E.$ and $Av. + P.E.$ being located, fifty per cent of the measures included in the distribution will be found between these two points. For the other two measures of variability the following statements may be made:

Between $Av. - \sigma$ and $Av. + \sigma$ there are 68.26 per cent of the measures in a distribution.

Between $Av. - A.D.$ and $Av. + A.D.$ there are 57.5 per cent of the measures included in the distribution.

It is possible to extend these statements with reference to these measures of variability. However, it is not usually done except for $P.E.$

Between $Av. - 2 P.E.$ and $Av. + 2 P.E.$ there are 82.26 per cent of the measures in the distribution. For a measure chosen at random from a distribution the chances are 4.5 to 1 that the measure chosen will lie between these two limits.

The chances that a measure chosen at random will lie between the $Av. + 3 P.E.$ and $Av. - 3 P.E.$ are 21 to 1. The per cent of measures included between these two limits is 95.70.

The chances that a measure chosen at random will lie between $Av. - 4 P.E.$ and $Av. + 4 P.E.$ are 142 to 1. The per cent of measures included between these limits is 99.30.

The chances that a measure chosen at random will lie between $Av. - 5 P.E.$ and $Av. + 5 P.E.$ are 1310 to 1. The per cent of measures included between these limits is 99.924.

The chances that a measure chosen at random will fall between $Av. - 6 P.E.$ and $Av. + 6 P.E.$ are 19,200 to 1. The per cent of measures falling between these two limits is 99.995.

The effect of chance errors of measurement upon the central tendency and measures of variability. Measures involve two types of errors—constant and variable. Constant errors affect central tendencies by making them larger or smaller by the amount of the error. If the constant error becomes known, corrections can easily be made in the central tendencies. Constant errors do not in any way affect measures of variability.¹ The effect of variable errors of measurement on central tendencies is decreased as the number of measures is increased. The

variable error of an average equals $\frac{1}{\sqrt{N}}$ times the variable

error measurement. Thus in order to reduce the variable error of the average by one half it is necessary to increase the number of measures four times.

Measures of variability are increased by the presence of variable errors. That is, a distribution of measures which include variable errors is spread out more than the distribution would be if the variable errors were eliminated. The effect of variable errors upon the standard deviation is given by the following equation:²

$$\sigma_{\text{True scores}} = \sigma_{\text{Obtained scores}} \sqrt{r.}$$

¹ Constant errors may be relative or absolute. A relatively constant error is one which bears a constant ratio to the measure. This type of constant error does affect measures of variability.

² Kelley, T. L. "The Measurement of Overlapping"; in *Journal of Educational Psychology*, vol. x, pp. 458-61 (November, 1919).

In this formula the true scores refer to the scores after the variable errors have been eliminated. The obtained scores include the variable errors. r is the coefficient of reliability.¹ Since both the probable error and the average deviation can be expressed in terms of the standard deviation, it is a simple matter to derive the corresponding formulæ for the other two measures of variability.

Effect of using a sample upon the central tendency and measures of variability. When a central tendency or measures of variability are interpreted as expressing generalizations, it is necessary to allow for the fact that they are calculated from a limited sample of measures. Assuming that the sample used was selected without bias, there is a chance that another sample selected in the same manner would yield a slightly different central tendency and measure of variability. It is, therefore, necessary to indicate the extent of the error which is introduced by this possibility. The formulæ generally used for this purpose are the following:

$$(1) \quad P.E._{average} = .6745 \frac{S.D._{distribution}}{\sqrt{N}}$$

$$(2) \quad P.E._{median} = .8454 \frac{S.D._{distribution}}{\sqrt{N}}$$

$$(3) \quad P.E._{standard deviation} = .6745 \frac{S.D._{distribution}}{\sqrt{2N}}$$

Comparison of two averages. In comparing two averages it is necessary to take into account the errors which each involve. There is a possibility of errors due to one or more of three causes — constant errors in the original measures, variable errors in the original measures, and sampling. The latter source does not apply when the averages are not

¹ See page 238.

treated as generalizations. The probable error of the difference of two averages due to variable errors of measurement and due to sampling is given by the formula,

$$P.E. \text{ Difference} = \sqrt{P.E. \text{ average}_1^2 + P.E. \text{ average}_2^2}$$

This assumes that the measures whose averages are being compared are unrelated.¹

The use of measures of deviation. Measures of deviation or variability have a number of important uses in the construction and use of educational tests.² The standard deviation (σ) is also used in the calculation of coefficients of correlation, as described in the following chapter.

EXERCISES

Assemble the following tables of data in frequency distributions. Exercise care to form the frequency distribution which will give the most useful summary of the data.

1. The numbers below represent a random sampling of the salaries of Illinois principals taken from the Illinois School Directory for 1920-21.

\$1650	\$1700	\$2000	\$2000	\$2000	\$2500	\$2000	\$3000	\$2000
2000	1650	1600	2300	2800	3000	2600	2700	2000
2000	1700	2025	2500	2200	2000	1500	1350	2300
1000	3000	2600	2700	1260	1500	1600	1800	1400
2259	2800	2400	2000	4500	3200	1800	3037	1500
2000	3000	1600	2000	2500	1560	2000	4000	2700
1800	3200	2000	3000	2000	2500	2000	3300	3384
2750	2250	1700	900	2400	1800	3000	2000	3000
3000	1700	1650						

2. The following are scores derived from Monroe's General Survey Arithmetic Scale II, Form 1, Grade VIII.

64	113	37	78	65	54	52	59	73	82	50	67
82	54	91	66	77	85	76	48	61	55	59	73
73	56	54	42	44	60	69	58	68	88	45	45
81	67	60	87	66	61	66	88	58	75	67	69
44	74	52	50	46	70	54	60	61	30	50	59
45	66	60	64	117	85	92	59	61	83	55	61
73	47	66	93	62	58	58	66	50	75	27	53
62	56	61									

¹ See page 209.

² See pages 95, 297, for descriptions of a number of these uses.

332 THEORY OF EDUCATIONAL MEASUREMENTS

3. Rate scores derived from Monroe's Standardized Silent Reading Test II, Form 2, Grade VII.

133	141	101	115	115	101	101	162	133	101	152
133	162	141	101	162	115	141	162	133	162	162
73	73	152	133	162	162	162	162	162	162	141
141	162	152	162	115	115	73	101	133	141	162
162	162	141	162	115	73	89	133	162	115	162
162	162	162	152	133	162	101	162	133	162	152
141	162	133	162	133	141	152	141	133	162	162
162	133	162	162	162	162	133	162	141	141	162
133	162	115								

4. The following numbers are intelligence quotients. They represent a random sampling of a small city school.

94	79	102	106	102	105	96	100	84	100	74
75	103	131	109	102	105	97	101	94	103	123
105	108	98	105	115	114	114	104	100	110	105
103	119	101	124	87	97	90	110	110	101	130
101	130	101	108	79	92	79	105	93	103	83
92	104	125	105	113						

5. Accuracy scores yielded by the Curtis Standard Research Tests in Arithmetic, Series B. Division.

89	88	71	67	0	91	100	86	88	88	71	62	67	78	67
73	100	78	80	0	44	100	100	100	75	83	100	78	90	100
50	73	67	100	80	57	57	85	89	87	91	90	85	75	100
87	100	20	83	89	50	82	60	87	89	87	89	25	100	91
87	78	83	100	100	44	92	50	100	83	73				

6. The numbers below represent the values of compositions written by high-school pupils as determined by the Nassau County Supplement to the Hillegas Scale. In rating the compositions each was assigned the numerical value of the scale composition which it was considered to equal most nearly in merit. Thus, a composition which is given a value of 7.2 is considered to be more nearly equivalent to the scale composition having this value than it is to any other step of the scale.

5.0	7.2	6.0	6.0	5.0	3.8	8.0	9.0	1.9	6.0	5.0	7.2	6.0
3.8	8.0	6.0	5.0	7.2	3.8	2.8	6.0	5.0	7.2			

Calculate the median, average, and standard deviation of the following frequency distributions:

7. Distribution of scores obtained from one algebra test. The score is the number of exercises solved correctly.

<i>Number solved</i>	<i>Number of pupils</i>
10.....	84
9.....	128
8.....	130
7.....	152
6.....	144
5.....	118
4.....	56
3.....	70
2.....	34
1.....	12
0.....	10
Number tested.....	938

8. Mental ages of pupils tested with Illinois Examination, Nov. 1920.

<i>M.A.</i>	<i>Number of pupils</i>
22.....	1
21.....	0
20.....	2
10.....	3
18.....	13
17.....	26
16.....	62
15.....	146
14.....	212
13.....	327
12.....	431
11.....	486
10.....	481
9.....	480
8.....	496
7.....	367
6.....	248
5.....	26

9. Height of male students, Ohio State University. (61 means nearer 61 than 60 or 62, etc.)

<i>Inches</i>	<i>Number of students</i>
74.....	4
73.....	9
72.....	23
71.....	75
70.....	87
69.....	109
68.....	126
67.....	106
66.....	93
65.....	57
64.....	38
63.....	11
62.....	10
61.....	2

10. Scores derived from Monroe's General Survey Arithmetic Scale II, Form 1, Grade VIII.

<i>Interval</i>	<i>Frequency</i>
120.....	0
110.....	2
100.....	0
90.....	3
80.....	9
70.....	11
60.....	27
50.....	23
40.....	9
30.....	2
20.....	1
Total.....	87

334 THEORY OF EDUCATIONAL MEASUREMENTS

11. Distribution of a random sampling of salaries paid to principals in Illinois.

<i>Interval</i>	<i>Frequency</i>
4500.....	1
4250.....	
4000.....	1
3750.....	
3500.....	
3250.....	3
3000.....	10
2750.....	3
2500.....	9
2250.....	5
2000.....	20
1750.....	4
1500.....	14
1250.....	3
1000.....	1
750.....	1
Total.....	75

12. Distribution of intelligence quotients. They represent a random sampling of a small city school.

<i>Interval</i>	<i>Frequency</i>
135.....	0
130.....	3
125.....	1
120.....	2
115.....	2
110.....	6
105.....	11
100.....	17
95.....	4
90.....	6
85.....	1
80.....	2
75.....	4
70.....	1
Total.....	60

13. Accuracy scores yielded by the Courtis Standard Research Tests in Arithmetic, Series B. Division.

<i>Interval</i>	<i>Frequency</i>
100.....	14
90.....	6
80.....	23
70.....	11
60.....	6
50.....	5
40.....	2
30.....	0
20.....	2
10.....	0
0.....	2
Total.....	71

14. Arithmetic Achievement Ages of an Illinois County. Illinois Exam. Oct. 1920.

<i>A.A.</i>	<i>Number of pupils</i>
17-6.....	1
17-0.....	
16-6.....	
16-0.....	
15-6.....	
15-0.....	
14-6.....	
14-0.....	3
13-6.....	1
13-0.....	8
12-6.....	3
12-0.....	4
11-6.....	12
11-0.....	15
10-6.....	12
10-0.....	36

14 *continued*

<i>A.A.</i>	<i>Number of pupils</i>
9-6.....	34
9-0.....	84
8-6.....	89
8-0.....	169
7-6.....	237
7-0.....	252
6-6.....	149
6-0.....	34
5-6.....	
5-0.....	
4-6.....	3
4-0-4-5.....	1
Total.....	1147

15. Scores derived from Buckingham's Scale for Problems in Arith. Form 1.

<i>Score</i>	<i>Number of pupils</i>
9.0-.....	328
8.5-.....	782
8.0-.....	1349
7.5-.....	2931
7.0-.....	58
6.5-.....	
6.0-.....	
5.5-.....	
5.0-.....	
4.5-.....	
4.0-.....	
3.5-.....	
3.0-.....	
2.5-.....	
0-2.4.....	1184
Total.....	6632

CHAPTER XIII

STATISTICAL METHODS: RELATIONSHIP EXISTING BETWEEN SETS OF PAIRED FACTS

Sources of paired facts. Paired facts are secured when two tests are given to the same pupils. This would occur when an arithmetic test and a silent reading test were given. We also have paired facts when a single test yields two scores, as, for example, a rate score and a comprehension score in the case of silent reading. Any two measures of the same pupil constitute a pair of facts. We may thus have paired facts resulting from a single test which has two scores, from two different tests, or from two applications of the same test. A test score may also be compared with the pupil's chronological age, his grade placement, his school marks, or with any other measure of his mental or physical traits. There are other sources of paired facts, but these are sufficient to indicate the need for methods of studying the relationship existing between them.

Arrangement of paired facts. In studying the relationship which exists between paired facts it is desirable to arrange them in a way which will assist in understanding the nature of this relationship. In general, the best arrangement is secured by tabulating the facts in the form of a correlation table. Sometimes this table is spoken of as a double-entry table. The character of the table is illustrated in Table XXIV.

Constructing a correlation table. The first step in constructing a correlation table is to decide upon the class intervals. The directions given on page 303 apply here. It

is not necessary that the number of intervals be the same for both sets of facts. Unless the distributions of the two sets of facts are irregular, the choice of intervals will affect the coefficient of correlation very little. The procedure to be followed in tabulating paired facts in a correlation table depends in part upon the form in which the original data are available. In most investigations in which correlation is used it is desirable to have each pair of facts recorded on a separate card. When this has been done, the best method of procedure is to sort the cards according to the intervals for one set of facts in the same way as they would be sorted if a frequency distribution was being formed. After the cards have been thus sorted each pile should be taken up in turn and sorted according to the other set of facts. The number of cards in each pile should then be recorded in the correlation table, as shown in Table XXIV, which is for rate scores obtained from Form 1 and Form 3 of the Courtis Silent Reading Test No. 2.

When all entries have been made the totals of the columns should be recorded in the line marked *f* at the bottom of the table. The totals of the lines should be entered in the column marked *f* at the right of the table. This column at the right of the table contains the distribution of the Form 1 scores. The line *f* at the bottom contains the distribution of the Form 3 scores. The totals of these two distributions should be identical. In Table XXIV they are 80.

Calculation of the coefficient of correlation by the Pearsonian formula. The processes of calculating the coefficient of correlation by means of the Pearsonian formula may be described in terms of the following steps:

Step I. Construct a correlation table as described above.¹

¹ The coefficient of correlation may be calculated without the construction of a correlation table. Two methods have been proposed for doing this. The most commonly used method is described in Rugg's

TABLE XXIV. CORRELATION TABLE FOR RATE SCORES OBTAINED FROM FORMS 1
AND 3 OF THE COURTIS SILENT-READING TESTS NO. 2, GRADE IV

Form 1

	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	f
30	1																										1
40		1																									
50			1																								
60				1																							
70					1																						
80						1																					2
90							1																				5
100								1																			3
110									1																		4
120										1																	7
130											1																10
140												1															3
150													1														9
160														1													2
170															1												8
180																1											3
190																	1										5
200																		1									3
210																			1								1
220																				1							2
230																					1						4
240																						1					1
250																											
260																											
270																											
280																											
380																											
f	1			1	2	2	6	5	5	5	8	8	10	1	6	3	7	3	1	1		2	1		1	1	80

Form 3

$$r = .85 \pm .01$$

Step II. Estimate the averages of the distributions at the mid-point of a convenient interval. Draw horizontal and

FORM 1

	0	5	10	15	20	25	30	35	40	45	50	55	60	65	f	d	fd	fd ²	Σxy	
65													1	1	1	3	6	18	108	90
60										1		1			2	5	10	50	30	
55								1			1	1			3	4	12	48	28	
50										2	3				5	3	15	45	39	
45								1	2	3	2	1	1		10	2	20	40	26	
40						2	1	4	3	3					13	1	13	13	2	
35			1	1	2	3	3	2		1	1				14	0	88			

FORM 2

30							3	1							4	-1	-4	4	3
25				1	3	2	2								8	-2	-16	32	38
20				3		1									4	-3	-12	36	42
15			3	2			1								6	-4	-24	96	96
10			2												3	-5	-15	75	60
5		1													1	-6	-6	36	36
0																			
f	1	6	7	5	9	11	10	6	9	6	4	1	1	76					
d	6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6						
fd	6	-30	-28	-15	-18	-11	108	6	18	18	16	5	6	619					
fd ²	36	150	112	45	36	11		6	36	54	64	25	36	611					

and at the bottom to provide for the following quantities: deviation from the assumed average (d), product of each deviation by its frequency (fd), and the product of the frequency and the square of the deviation (fd^2). This procedure is identical with the calculation of the standard deviations of the two frequency distributions. The standard deviation of the horizontal distribution is called σ_x and the standard deviation of the vertical distribution is called σ_y .

Step III. The two columns at the extreme right are headed Σxy . The entries to this heading may be either positive or negative. It is customary to provide space for two columns, one for the positive entries and the other for the negative entries. The procedure for determining the entries in these columns is as follows: Each frequency in the correlation table is multiplied by the product of its deviations from both assumed averages. In Fig. 12 there is a frequency of 1 in the first line under 55. The deviation of this from the assumed average of the vertical distribution is 6. The deviation of this from the assumed average of the horizontal distribution is 4. This frequency is, therefore, to be multiplied by the product of 6 and 4 or 24.

Since all of the frequencies in a row have the same deviations from the assumed average of the vertical distribution, the arithmetical work will be reduced if each frequency in a row is multiplied by its deviation from the assumed average of the horizontal distribution, and the sum of these products multiplied by the deviation of the row from the assumed average of the vertical distribution. In Fig. 12 the sum of the products of the frequencies and their deviations from the assumed average of the horizontal distribution for the "65 row" is $(4 \times 1) + (5 \times 1) + (6 \times 1) = 15$. This multiplied by 6 gives the total amount of 90, which is the entry to be made in the Σxy column. The other entries in this column are determined in a like manner.

It is necessary to bear in mind the sign of the deviation. For example, the entries on the "40 line" are partly to the left of the assumed average and partly to the right. The first frequency is 2. Its deviation is -2 . The product is -4 . The next frequency is 1 with a deviation of -1 . This gives a product of -1 . The frequency 4 has a deviation of 0, hence the product is 0. The first 3 has a deviation of 1, and the second 3 a deviation of 2. The sum of these products is $-4 - 1 + 3 + 6 = 4$. This, multiplied by 1, the deviation of the row from the assumed average of the vertical distribution, gives the product of 4 which is entered in the last column of the figure. Sometimes the product will be negative. In such cases the entries are made in the "minus" column.

Step IV. The coefficient of correlation is now calculated by the formula:

$$r = \frac{\frac{\sum xy}{N} - c_x c_y}{\sigma_x \sigma_y}$$

In obtaining $\sum xy$ we take the algebraic sum of the products. In Fig. 12 all of them are positive. If some were negative the sum would be the difference between the positive and negative products. In Fig. 12 the product of $c_x c_y = -.0714$. Since this negative product is to be subtracted it means the addition of a positive quantity. The calculation of the coefficient of correlation is completely illustrated in Fig. 12. This arrangement of the work is recommended.

Interpretation of the coefficient of correlation. The general appearance of a correlation table suggests the type of relationship which exists between the two sets of quantities recorded in it. If all the entries fall upon the diagonal line, as is shown in Table XXV, there is a perfect relationship or correspondence between the two quantities. In this

illustration the two paired facts are the cost of an amount of coal and the number of tons, the price remaining constant. The table is to be read as follows: One lot of 2 tons costs \$20, four lots of 4 tons each cost \$40 each, eight lots of 6 tons each cost \$60 each, and so on. If we know the number of tons of coal, the cost of that amount is definitely determined. If we pass in one set of facts from small to larger quantities, the paired facts also increase regularly, and always in the same ratio. In case the entries fall upon the other diagonal line of the table there is a definite relationship between the two sets of facts. However, this relationship is inverse. When one fact is large the other is small. The coefficient of correlation is a numerical index of the relationship between the two sets of paired facts. When all the entries in the correlation table fall upon the diagonal line, as in Table XXV, the coefficient is $+1.00$. If the entries all fall upon the opposite line of the table the coefficient is -1.00 .

TABLE XXV. ILLUSTRATING PERFECT CORRELATION. THE
RELATION BETWEEN THE COST OF VARIOUS AMOUNTS OF
COAL AND THE AMOUNT

If we measure the weight and height of a number of children we have two sets of paired facts. For each child there are two facts — his height and weight. We know from general observation that, in general, tall children weigh more than short ones. We also know that there are numerous exceptions to this rule. Some relatively short children weigh more than others who are taller. We may describe the relation between the height and weight of children by saying that there is a tendency for tall children to weigh more than short ones, although this relationship is not perfect. Hence the coefficient of correlation for these two sets of facts will be less than 1.00.

If twelve dice are thrown one hundred times and the number of 4's is recorded for each throw, the two sets of paired facts formed by pairing together the odd numbered throws and the immediately following even numbered throws are not in any way related except by chance. For example, if an odd numbered throw resulted in eight 4's, no prediction can be made concerning the probable number of 4's which will appear on the next even numbered throw. There is absolutely no connection between these two sets of paired facts. If these paired facts were recorded in a correlation table they would be scattered over the entire table in a miscellaneous order. The coefficient of correlation for such a table would be approximately zero. Secrist ¹ gives a correlation table of 500 pairs of throws of twelve dice. The coefficient of correlation is $r = .04 \pm .03$.

Facts obtained in mental measurements do not exhibit perfect relationship except by chance. Correlation tables are not obtained in which all the entries fall upon a diagonal line. Frequently they are grouped near this line, but in other cases they are scattered over most of the correlation table.

¹ Secrist, Horace, *An Introduction to Statistical Methods*, p. 434. Macmillan, 1917.

A few concrete illustrations of correlation may be helpful. The coefficient of correlation of the height of fathers with the height of sons has been calculated to be .51. For the age of husband with the age of wife it is .91. The number of mothers' children with the number of daughters' children gives a coefficient of correlation of .21. The correlation between success in mathematics and success in English has been found to be .40. Reliability coefficients of most educational tests fall between .60 and .90.¹

Effect of variable errors and constant errors upon the coefficient of correlation. The presence of absolute constant errors in either one or both sets of paired facts has no effect whatever upon the coefficient of correlation. Relative constant errors may have a slight effect. The presence of variable errors always tends to lower the coefficient of correlation. When two measurements of each of the two traits which are being studied are available it is possible to correct for the effect of the presence of these errors by certain formulæ.² This is called correction for attenuation.

The effect of sampling upon the coefficient of correlation. In the study of paired facts we may seek to ascertain if any relation exists between the traits of which we have a limited number of measures. For example, we may ask if there is any relation between achievement in Latin and achievement in English, or achievement in arithmetic and achievement in silent reading. The usual procedure in studying such questions is to secure, for a number of representative pupils, measures of their achievements in the two fields under consideration. These measures may be the scores obtained from standardized tests, or they may be the school marks received in the two school subjects. The coefficient of corre-

¹ See page 203.

² Thorndike, E. L. *An Introduction to the Theory of Mental and Social Measurements*. Teachers College, Columbia University, 1916, p. 179.

lation is an index of the relation which exists between such sets of paired facts.

However, in interpreting it is necessary to bear in mind that a coefficient of correlation is necessarily computed from the facts gathered. These are generally only a sample of all of the existing facts. In a study of the relation of achievement in Latin to achievement in English, measurements of these achievements would be secured from only a few of the total number of pupils. The study would be unusually comprehensive if several hundred cases were included. We are, therefore, endeavoring to determine a general relationship upon the basis of a sample. For this reason it is necessary to make due allowance for the possibility of the sample not being completely representative of the total, even though it was chosen in an unbiased or random manner.

The addition of other groups of facts to our correlation table may change its general appearance, and the coefficient of correlation calculated from it. In fact, this will happen, unless the additional facts are related to each other in exactly the same way as the facts in our first table. Hence, the coefficient of correlation calculated from one group of measures of the traits would likely differ slightly from that calculated from another group of measures. The average of all such coefficients of correlation is taken as the true index of the relationship. The difference of any coefficient from this average is called the "error due to sampling." The magnitude of the error of a coefficient of correlation, due to sampling, may be described in terms of the limit which is exceeded by 50 per cent of the coefficients of correlation calculated from similar collections of these paired facts. This limit is the probable error.

Computing *P.E.* of the coefficient of correlation, due to sampling. The probable error of the coefficient of correla-

tion, due to sampling, can be computed by the following formula:¹

$$P.E._r = .6745 \frac{1 - r^2}{\sqrt{N}}$$

The meaning of the probable error of the coefficient of correlation may be illustrated as follows: If a coefficient of .20 has been obtained from a correlation table based on 36 cases; i.e., 36 pairs of facts, the probable error of this coefficient of correlation is .11. This means that the chances are 1 to 1 that the true coefficient lies between .09 and .31. The chances are 4.5 to 1 that it lies between — .02 and .42. The chances are 21 to 1 that it lies between — .13 and .53.

It is obvious that we are not justified in asserting that the true coefficient of correlation — i.e., the one calculated from *all* of the pairs of facts, or the average of the coefficients calculated from a large number of collections of 36 pairs of facts — is positive. Therefore, a coefficient of correlation of .20 with a probable error of $\pm .11$ cannot be considered proof that a relationship exists between the two traits. Although the calculated coefficient is positive, the probable error is so large in comparison with the coefficient of correlation that it is not altogether certain that the true coefficient would be positive. For this reason, it is necessary to require that the coefficient be several times larger than its probable error before it is interpreted as proving the existence of a relationship. Since the probable error of a coefficient of correlation is frequently needed in its interpretation, it is customary to express them together as $r = .20 \pm .11$.

Authorities differ with reference to the ratio which must exist between the coefficient of correlation and its probable error before the existence of a relationship can be asserted.

¹ Tables have been prepared from which the value of $P.E._r$ can be read for various values of r and N .

By at least one writer the ratio is placed at 6. Another writer places it as low as 2 or 3. A conservative rule is that the coefficient must be four times its probable error before the existence of a relationship can be assumed.

What a relationship between measures means. Proof that relationship exists between two quantities means only that they *tend* to change together. If one quantity becomes larger there is a tendency for the other also to become larger in case the relationship is direct. If it is inverse, as one quantity becomes larger there is a tendency for the other to become smaller. The larger the coefficient of correlation, the more marked is the relationship. For the relation of achievement of one school subject to that of another school subject, most of the coefficients of correlation range from .30 to .50. Thus, .60 is considered to indicate a high degree of correlation for these two traits. This only means that .60 is larger than most of the coefficients of correlation obtained in the study of the relationship existing between the achievement in different school subjects. In the study of the reliability of educational tests, coefficients of correlation from .60 to .90 are most frequent when the scores are secured from pupils belonging to a single grade. Hence, for this relationship, .90 is considered high and .60 is low.

The establishment of the existence of a relationship between two traits does not mean that one is the cause of the other. For example, there is a positive relationship between achievement in Latin and achievement in English, but that does not necessarily mean that success in Latin is the cause of success in English. It may be that success in both subjects is dependent upon a third factor, such as general intelligence. The interpretation of coefficients of correlation requires an intimate acquaintance with the traits which are being studied, and in many cases with other traits.

The departure from perfect correlation. In studying the

objectivity and reliability of educational tests we are not concerned with establishing the existence of a relationship between the two sets of scores. The presumption in favor of a relationship is so strong that its existence can in general be assumed. In most cases it is marked. Our interest is rather in knowing the amount of departure from complete or perfect correlation. The coefficient of correlation indicates this only in a very general way. It is, therefore, an unsatisfactory index of the objectivity and reliability of educational tests. The *probable error of estimate* is proposed as a more significant index. In order to understand its meaning it is necessary to consider first the regression equation.

The regression equation. The general relation between two sets of paired facts may be represented by the regression equations:

$$y = r \frac{\sigma_y}{\sigma_x} x \qquad x = r \frac{\sigma_x}{\sigma_y} y \qquad (A)$$

In these equations x and y are deviations from their respective averages, σ_x and σ_y are the standard deviations, and r is the coefficient of correlation existing between the two sets of facts. If y' represents the absolute magnitude of the y -facts, and x' the absolute magnitude of the x -facts, and a_x and a_y are the averages of the two distributions, we may write these equations in terms of them as follows:

$$(y' - a_y) = r \frac{\sigma_y}{\sigma_x} (x' - a_x) \qquad (x' - a_x) = r \frac{\sigma_x}{\sigma_y} (y' - a_y) \qquad (B)$$

It is obvious that unless $r = \pm 1.00$ there is no single relationship which holds for all pairs of facts. These equations express the best summary or "average" of the various relationships which exist. For Table XXIV, $r = .85 \pm .01$, $\sigma_x = 47.9$, $\sigma_y = 55.2$, $a_x = 150.0$ and $a_y = 154.5$. If these values are substituted in equations B, we have:

$$y' = .98x' + 7.5 \qquad x' = .74y' + 36$$

One interpretation of the equation $y' = .98x' + 7.5$ is that for every unit of change in x' there is in general a change of .98 unit in y' . Since in this illustration x' and y' are in terms of comparable units (words per minute) this shows that a close relationship exists between them. If for each x -fact there is estimated a corresponding y -fact by means of the equation $y' = .98x' + 7.5$, the coefficient of correlation of the estimated y -facts with the x -facts will be 1.00. This must be true because the pairs of facts lie on the straight line whose equation is $y' = .98x' + 7.5$. Any set of y -facts estimated from the x -facts by means of an equation of the form $y = mx + c$ would yield the same coefficient of correlation. The estimates made by means of the regression equation, $y' = .98x' + 7.5$, involve the least possible changes from the actual y -facts.¹ Therefore, the difference between actual y 's and the estimated y 's furnishes an indication of their departure from perfect correlation. The equation $x' = .74y' + 36$ may be used to obtain a set of estimated x -facts which will correlate perfectly with the y -facts.

In Table XXVI are given the differences between the actual score and the estimated score for a number of pupils whose scores are recorded in Table XXIV. The score on Form 1 of the Courtis Silent Reading Test No. 2 is the x -fact, and the score on Form 3 the y -fact. By means of the equation

$$y' = .98x' + 7.5$$

estimated scores on Form 3 have been calculated. These estimated Form 3 scores correlate perfectly with the Form 1 scores. The differences between the actual Form 3 scores

¹ The sum of the squares of the differences between the estimated y 's and the actual y 's is a minimum when the estimation is made by means of the equation $y = r \frac{\sigma_y}{\sigma_x} x$, or $(y' - a_y) = r \frac{\sigma_y}{\sigma_x} (x' - a_x)$. (Yule, G. U., *An Introduction to the Theory of Statistics*, pp. 171-72).

TABLE XXVI. SHOWING ACTUAL SCORES AND ESTIMATED FORM 3 SCORES ON THE COURTIS FORM 3 RATE SCORES ON THE COURTIS SILENT READING TESTS NO. 2

<i>Pupil No.</i>	<i>Score on Form 1</i>	<i>Score on Form 3</i>	<i>Est. score on Form 3</i>	<i>Diff.</i>	
				+	-
1	98	100	104		4
2	38	34	45		11
3	130	126	135		9
4	203	187	206		19
5	98	83	104		21
6	69	70	75		5
7	94	110	100	10	
8	150	148	155		7
9	85	89	91		2
10	97	93	103		10
11	157	160	161		1
12	106	107	111		4
13	148	162	153	9	
14	130	141	135	6	
15	128	152	133	19	
16	182	180	186		6
17	157	149	161		12
18	116	130	121	9	
19	108	113	113	0	
20	93	127	99	28	
21	81	72	87		15
22	124	120	129		9
23	154	188	158	30	
24	182	198	186	12	
25	113	127	118	9	

and the estimated Form 3 scores are given in the last column of the table. These differences indicate how much the scores obtained from the two forms of the tests fall short of perfect correlation. For the first pupil, the difference is - 4. For the second, it is - 11. The differences vary. Many are relatively small, but others are large. They are about equally divided as to sign. They represent the departures from perfect correlation.

The probable error of estimate. It is possible to describe these differences of estimated scores from the actual scores

without calculating them. If y is taken to represent the actual score, $r \frac{\sigma_y}{\sigma_x} x$ is the obtained score.¹ The difference is

$$y - r \frac{\sigma_y}{\sigma_x} x.$$

If the correlation is rectilinear and the distributions are normal, these differences will form a normal distribution with the average at zero. Hence the standard deviation of the distribution of the differences or departures from perfect correlation may be expressed as follows:

$$\begin{aligned} (\sigma_{Est_y})^2 &= \frac{1}{N} \sum (y - r \frac{\sigma_y}{\sigma_x} x)^2 \\ &= \frac{\sum y^2}{N} - 2r \frac{\sigma_y}{\sigma_x} \frac{\sum xy}{N} + r^2 \left(\frac{\sigma_y}{\sigma_x} \right)^2 \frac{\sum x^2}{N} \\ &= \sigma_y^2 - 2r \sigma_y^2 \frac{\sum xy}{N \sigma_x \sigma_y} + r^2 \sigma_y^2 \\ &= \sigma_y^2 - r^2 \sigma_y^2 \\ &= \sigma_y^2 (1 - r^2) \quad \text{or,} \\ \sigma_{Est_y} &= \sigma_y \sqrt{1 - r^2} \end{aligned}$$

By the same method we can obtain

$$\sigma_{Est_x} = \sigma_x \sqrt{1 - r^2}$$

Since the probable error permits of a more convenient interpretation than the standard deviation, we generally use the following equations and obtain the *probable error of estimate*. These expressions give the median deviation of the departures from perfect correlation.

¹ In this proof x and y are taken as deviations from the true averages. The absolute values could be used but to do so would make the proof cumbersome.

$$P.E._{Est_x} = .6745 \sigma_x \sqrt{1 - r^2}$$

$$P.E._{Est_y} = .6745 \sigma_y \sqrt{1 - r^2}$$

Apply these formulas to Table XXIV.

$$\begin{aligned} P.E._{Est_x} &= (.6745) (47.9) \sqrt{1 - .85^2} \\ &= 16.79 \end{aligned}$$

$$\begin{aligned} P.E._{Est_y} &= (.6745) (55.2) \sqrt{1 - .85^2} \\ &= 19.35 \end{aligned}$$

This means that half of the scores obtained from Form 3 of this test failed of perfect correlation with the Form 1 scores by 19.35. For any pupil selected at random from this group the chances are 1 to 1 that his Form 3 score would depart from perfect correlation with his Form 1 score by as much as 19.35 words per minute. A similar interpretation may be given for the departure of the Form 1 scores from perfect correlation with the Form 3 scores. Since the two forms are assumed to be equivalent, the differences of the two probable errors of estimate may be considered to be due to chance variations in the values of x and y . Hence, the average of the two determinations should be taken as the most representative of the Courtis Silent Reading Test No. 2.

$$P.E._{Est} = 18.07$$

The error of estimate represented graphically. Fig. 13 illustrates the meaning of the error of estimate. The scatter diagram is for the scores for Form 1 and Form 2 of the Illinois General Intelligence Scale, when they were given to a group of fifth-grade pupils. The coördinates of each dot represent the two scores of a pupil. The oblique line represents the regression equation, $y = 4.92 + .80x$. Perfect correlation of the Form 1 scores with the Form 2 scores would be secured if each dot in the figure were shifted ver-

tically so that it would fall upon this regression line. A few short vertical lines have been drawn in the figure to indicate the magnitude of the shifts required. Some are very small,

FORM 1

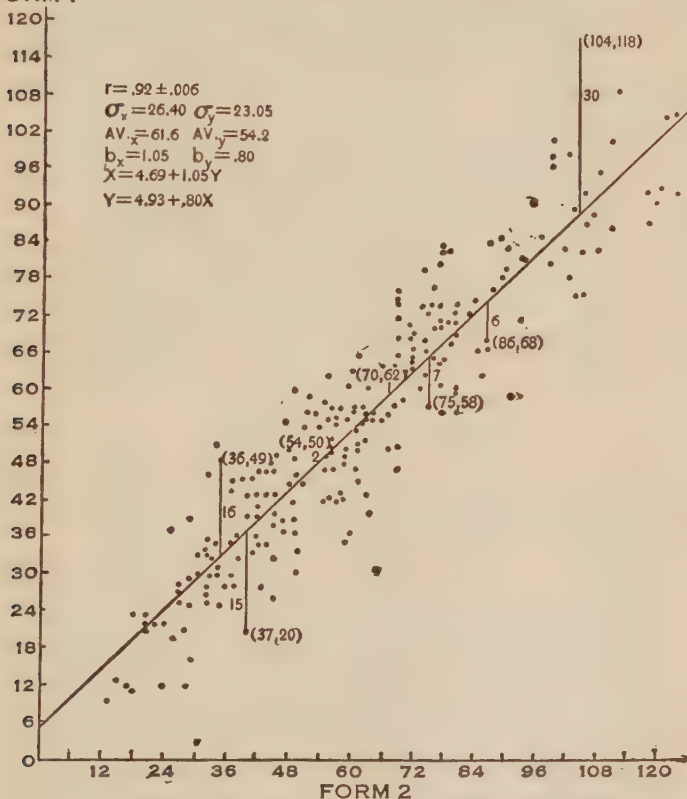


FIG. 13. CORRELATION OF FORM 1 SCORES WITH FORM 2 SCORES OF THE ILLINOIS GENERAL INTELLIGENCE SCALE, FIFTH GRADE

others are relatively large. Some are positive, others are negative. The probable error of estimate is the median amount of shift that is required to secure perfect correlation.

Since we are interested in how nearly our paired facts approach perfect correlation, we may describe the condition which exists in terms of the departure from perfect correlation. From this point of view Fig. 13 represents the extent to which the Form 1 scores depart from perfect correlation with the Form 2 scores. The probable error of estimate is thus a measure of the departure from perfect correlation.

Probable error of measurement. When the paired facts, whose relation is being studied, are the scores derived from two applications of the same test or of duplicate forms of a test, it is possible to derive a formula which gives us a measure of the departure of either set of scores from perfect correlation with a corresponding set of true scores. The derivation of this formula has been explained on page p. 214. It is as follows:

$$P.E.M = \frac{\sigma_1 + \sigma_2}{2} \sqrt{1 - r_{12}}$$

When one is studying the reliability of a test this formula is recommended in preference to the one for the probable error of estimate.

Coefficients of probable errors of estimate and probable errors of measurement. By means of the formulæ given above for the probable error of estimate and the probable error of measurement, we may calculate for any value of r_{12} the factors by which the standard deviation is to be multiplied in obtaining these measures of departure from perfect correlation. In Table XXVII, the coefficients of the probable error of estimate and the probable error of measurement are given for values of r_{12} ranging from .00 up to .99. To find either the probable error of estimate or the probable error of measurement for a given value of r_{12} , it is only necessary to multiply the standard deviation by the appropriate coefficient.

This table helps us to realize the significance of a coefficient of correlation. A coefficient of correlation of .65 is considered high, and yet we find by means of this table that

TABLE XXVII. COEFFICIENTS OF PROBABLE ERRORS OF ESTIMATE AND PROBABLE ERRORS OF MEASUREMENT

<i>Coefficient of correlation, r_{12}</i>	<i>Probable error of estimate</i>	<i>Probable error of measurement</i>
.00	.6745	.6745
.05	.6736	.6574
.10	.6711	.6399
.15	.6669	.5544
.20	.6609	.6033
.25	.6531	.5841
.30	.6434	.5644
.35	.6306	.5438
.40	.6182	.5225
.45	.6023	.5002
.50	.5841	.4769
.55	.5634	.4525
.60	.5396	.4266
.65	.5126	.3990
.70	.4817	.3694
.75	.4461	.3373
.80	.4047	.3016
.85	.3553	.2612
.90	.2940	.2132
.92	.2643	.1907
.94	.2301	.1652
.95	.2106	.1508
.96	.1889	.1349
.97	.1638	.1168
.98	.1342	.0952
.99	.0952	.0674

in the case of 50 per cent of the pairs of values, there is a departure from perfect correlation which equals or exceeds .5126 of the standard deviation of the set of measures being considered. In fact, for coefficients of correlation of .90 or higher, both the probable error of estimate and the probable error of measurement are relatively large. Thus, from

the standpoint of departures from perfect correlation, it is clear that values of r_{12} which we have been accustomed to interpret as indicating a "high correlation" are really indicative of departures from perfect correlation so large that it becomes necessary to revise our concept of "high correlation."

Effect of selection of data upon coefficients of correlation.

We have already noted that, since it is necessary in studying general relationships to use limited samples of data, it is necessary to take into account the probable error of the coefficient of correlation due to sampling. In the case of data collected from school children it is necessary to bear in mind, in addition to the above, the type of the population from which the data are collected. We may collect data from pupils belonging to a single school grade, or to a series of school grades. In Table XXVIII coefficients of correlation for pupils in grades III-B to VIII-A of one school system are given. In every case, all children of the school system were tested. Each grade group is sufficiently large so that the probable error of the coefficient of correlation, due to sampling, is relatively small. In the third column, the coefficients of correlation of mental age with chronological age are given. For each grade group they are negative. Several of them are sufficiently large negative quantities to indicate a distinct inverse relationship between chronological age and mental age within the population group from which these data were secured. This, of course, is in agreement with our general observations. The bright pupils in any grade are the younger ones. The older pupils in any grade are generally the dull ones. However, when all grades are combined, the coefficient of correlation of chronological age with mental age is .56, which indicates a distinct positive relationship between mental age and chronological age when all grades are taken together. This again is in accord

with our general observations. Hence, we have an illustration which emphasizes that the coefficient of correlation is materially affected by the grade-range of the population group from which the data are collected.

TABLE XXVIII. EFFECT OF NATURE OF TRAIT AND SELECTION OF POPULATION GROUP UPON COEFFICIENT OF CORRELATION

Grade	No. of pupils	Chronological Age and Mental age	Achievement Quotient and Intelligence Quotient			Chronological Age and Achievement Age			Mental Age and Achievement Age		
			Arith.	Reading		Arith.	Reading		Arith.	Reading	
				Comp.	Rate		Comp.	Rate		Comp.	Rate
3B	314	-.12	-.62	-.04	-.18	-.10	-.21	-.06	.29	.63	.24
3A	268	-.20	-.62	-.14	-.21	-.04	-.22	-.05	-.12	.48	.02
4B	449	-.30	-.62	-.18	-.12	-.10	-.26	-.04	.52	.59	.28
4A	241	-.35	-.36	-.02	-.23	-.33	-.33	-.26	.38	.46	.31
5B	454	-.31	-.37	-.13	-.19	-.33	-.23	-.19	.47	.59	.39
5A	255	-.40	-.19	-.01	-.01	-.33	-.38	-.27	.44	.63	.50
6B	426	-.44	-.39	-.01	-.08	-.41	-.31	-.18	.57	.64	.45
6A	224	-.40	-.60	-.16	-.20	-.32	-.28	-.14	.44	.54	.28
7B	399	-.31	-.43	-.19	-.07	-.06	-.19	-.16	.43	.50	.34
7A	201	-.33	-.45	-.17	-.01	-.20	-.28	-.20	.27	.67	.39
8B	372	-.40	-.36	-.06	-.02	-.20	-.16	-.15	.51	.61	.44
8A	214	-.37	-.34	-.15	-.12	-.28	-.34	-.23	.44	.60	.51
All Grades		.56	-.41	-.10	-.14	.51	.38	.25	.76	.75	.40

In the next three columns the coefficients of correlation for intelligence quotients with achievement quotients are given. These coefficients are negative. In the case of arithmetic there are relatively large negative quantities. In fact, they are slightly larger than the ones for chronological age with mental age. When all grades are combined the coefficients of correlation are all negative, and differ but little from the average of the coefficients of correlation for the separate grade groups. Here we have a distinct reversal of the condition illustrated in Column 3. Hence it appears that the coefficient of correlation is affected not only by the

grade-range of the population group, but also by the nature of the traits which are considered.

In the case of both the intelligence quotient and the achievement quotient the average is practically the same for all grade groups. The standard deviations are also essentially the same. Not only are they the same for all grade groups, but when all grades are combined approximately the same averages and standard deviations are secured. We are in this case dealing with traits which do not increase from grade to grade, but tend to remain essentially constant. In the case of chronological age and mental age the traits increased from grade to grade. This is the explanation of the apparent inconsistency between the results for the two groups of traits.

In the remaining columns of the table other coefficients of correlation are given which further illustrate the fact that, in interpreting coefficients of correlation, it is necessary to bear in mind the character of the population group from which the data were collected, and also the nature of the traits measured.

EXERCISES

Form appropriate correlation tables for the following sets of paired facts. Then calculate the coefficient of correlation and the probable error of estimate.

1. RATES OF SILENT READING MEASURED BY FORM 1 AND FORM 3 OF
COURTIS SILENT READING TEST NO. 2

<i>Form 1</i>	<i>Form 3</i>	<i>Form 1</i>	<i>Form 3</i>	<i>Form 1</i>	<i>Form 3</i>	<i>Form 1</i>	<i>Form 3</i>
159	169	118	100	102	112	128	152
113	127	106	96	118	121	135	143
146	130	203	187	154	183	98	100
193	223	193	191	130	141	150	148
188	198	193	213	141	121	148	139
142	163	157	143	144	149	103	99
175	181	153	163	146	120	98	83
85	89	157	160	182	180	126	117
93	127	108	113	157	149	173	387
157	207	139	129	193	147	124	120
205	186	174	162	120	113	120	123
118	108	175	208	183	198	116	130
77	80	151	148	69	70	148	152
280	286	130	124	193	230	148	162
160	162	134	148	214	223	193	177
93	81	279	289	254	236	153	169
247	241	81	72	97	93	206	174
94	110	38	34	131	113	177	176

2. SCORES ON PRESSEY PRIMER TEST AND INDIANA SCALE OF ATTAIN-
MENT NO. 1

<i>P.P.</i>	<i>Ind. Att.</i>	<i>P.P.</i>	<i>Ind. Att.</i>	<i>P.P.</i>	<i>Ind. Att.</i>	<i>P.P.</i>	<i>Ind. Att.</i>
65	21	59	21	62	14	29	9
69	17	71	20	55	8	51	9
50	11	27	20	61	18	43	12
52	15	50	22	68	10	70	22
50	22	52	20	34	19	6	4
55	22	79	22	73	20	54	19
60	21	45	14	36	10	53	14
30	22	20	22	18	3	59	16
13	22	30	21	57	21	54	6
67	21	51	21	41	17	60	17
69	14	56	6	38	11	8	8
83	20	56	21	25	16	49	16
65	16	39	17	43	17	19	6
66	22	58	20	45	18	84	18
82	15	63	21	18	4	41	12

3. CHRONOLOGICAL AGE AND INTELLIGENCE SCORE. GRADE 3A

<i>Chron. age</i>	<i>Intelli- gence score</i>	<i>Chron. age</i>	<i>Intelli- gence score</i>	<i>Chron. age</i>	<i>Intelli- gence score</i>	<i>Chron. age</i>	<i>Intelli- gence score</i>
9	28	9	24	10	32	10	12
10	33	10	36	9	55	10	22
9	31	9	27	10	25	10	35
9	45	9	34	9	44	9	44
9	38	15	27	11	43	9	52
9	47	10	31	10	37	9	31
9	48	9	28	10	49	9	40
10	73	8	43	9	56	9	38

Calculate the coefficient of correlation and the probable error of estimate for the following correlation tables.

4. CORRELATION TABLE FOR SCORES YIELDED BY SERIES 2A AND SERIES 2B OF BARR'S DIAGNOSTIC TEST IN AMERICAN HISTORY

Series 2B	Series 2A															
	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	T
80													1			1
75												1	1		1	4
70							2	1	2	1	1	3				10
65							1	3	2	6	4		2	1		17
60						1	2	8	6	2	1			1		21
55					2	1	1	6	6	3		1				20
50					4	4	4	1	3	1						17
45					7	1	5		2	1		1				17
40				1	3	8			1							13
35				2			1									3
30				1	1		1		1							4
25	1															1
T	1			4	17	15	17	19	21	15	6	6	4	2	1	128

5. CORRELATION TABLE FOR SCORES YIELDED BY FORM 1 AND FORM 2 OF THE ILLINOIS GENERAL INTELLIGENCE SCALE

Form 2	Form 1													T
	5	10	15	20	25	30	35	40	45	50	55	60	65	
65											1	1	1	3
60									1		1			2
55							1			1	1			3
50									2	3				5
45						1	2	3	2	1	1			10
40					2	1	4	3	3					13
35		1	1	2	3	3	2		1	1				14
30						3	1							4
25			1	3	2	2								8
20			3		1									4
15		3	2			1								6
10		2			1									3
5	1													1
T	1	6	7	5	9	11	10	6	9	6	4	1	1	76

6. CORRELATION TABLE FOR CHRONOLOGICAL AGE AND MENTAL AGE OF SIXTH-GRADE PUPILS

Mental age	Chronological age																T
	9	9½	10	10½	11	11½	12	12½	13	13½	14	14½	15	15½	16		
18					2											2	
17																	
16																	
15		2	1	4	5	2										14	
14	1	1		3	4	4										13	
13		1	3	4	13	6	6	4		1		1				39	
12		2	4	10	27	16	8	4	1	1		1		1		75	
11		2	5	9	26	21	14	8	5	1						91	
10		1	3	6	30	27	10	7	9	3	2		2			100	
9				3	8	13	10	8	6		1	1		1	1	51	
8					2	5	5	10	1	1	4	1	1			32	
7						1	1	1	2	2						2	
6										1		1				7	
T	1	9	16	41	117	95	54	42	24	10	7	5	3	2		426	

INDEX

- Ability, nature of, 66; relation to performance, 66, 193; relation to score, 194.
Accuracy, 108 ff.; relation to difficulty, 119.
Achievement age, 155.
Achievement quotient, 157, 180, 246, 268.
Achievement tests, 40.
Alexander, Carter, 244.
Ashbaugh, E. J., 189.
Assumptions, 21.
Average, 309.
Average deviation, 322.
Averages, comparison of, 330.
Ayres, L. P., 8, 78, 90, 142.
- Binet, A., 3, 9.
Brooks, F. S., 116.
Brown, Wm., 205.
Buckingham, B. R., 9, 152.
Bureau of Educational Research, 13.
Burgess, May Ayres, 87, 246.
- Charters, W. W., 90, 118, 288.
Chronological age norms, 161.
Classification of pupils, 43.
Coaching, effect of, 168.
Coefficient of correlation, interpretation of, 341.
Coefficient of correspondence, 218.
Coefficient of reliability, 202.
Combined scores, 130.
Common unit, 147.
Comparable measures, 211.
Completion examination, 286.
Composite scores, 224.
Constant errors, 198, 243, 344.
Continuous series, 301, 314.
Cost of testing, 235.
Courtis, S. A., 7, 12, 24, 186, 219, 264, 269.
Criterion measures, 221.
Cycle test, 62, 75.
- Diagnosis, 46, 177, 245, 249.
Diagnostic tests, 40.
Difficulties, 61, 94, 119.
Discrete series, 302, 313.
Discrimination, 219.
Duplicate forms, 169.
- Educational quotient, 157.
Efficiency, 44, 172, 264.
Elliot, E. C., 28.
Errors, see constant errors and variable errors.
Examination marks, 28.
Examinations, increasing objectivity of, 282; norms for, 290, 294.
Exercise, 56, 76; selection of, 89; weighting of, 116.
- Fordyce, Chas., 116.
Franzen, Raymond, 157.
Freeman, F. N., 142.
Frequency distribution, 299, 303.
Function, 186.
- General intelligence, nature of, 39.
Grade norms, 161, 179, 245.
Gray, C. T., 133.
Gray, W. S., 130.
- Henmon, V. A. C., 145.
Hotz, H. G., 127, 145.
- Index of reliability, 206.
Intelligence tests, group, 10.
Inter-grade interval, 98.
Irregular test, 62, 75, 108.
- Johnson, F. W., 31.
- Kelley, T. L., 117, 123, 195, 206, 211.
Kelly, F. J., 11, 30, 116, 280.
- Law of single variable, 87.

- McCall, W. A., 148, 150, 155, 157, 284.
 Measurement defined, 15, 18.
 Median, 313.
 Mental age norms, 161, 179.
 Minnick, J. H., 117.
 Mode, 321.
 Monroe, Walter S., 78, 116, 156, 170, 203, 289.
 Murdoch, Katharine, 134, 139, 141.

 New examination, 286.
 Normal distribution, 306.
 Norms, basis of, 162.

 Objective, 26.
 Objectivity, 196.
 Otis, A. S., 213.

 Percentile, 319.
 Percentile scores, 154.
 Performance, variability of, 189.
 Pintner, Rudolph, 152, 155.
 Point scores translated into school marks, 292.
 Power test, 63.
 Practice effect, 167.
 Pressy, L. W., 222.
 Pressy, S. L., 222.
 Probable error, 327.
 Probable error of estimate, 348.
 Probable error of measurement, 207, 354.
 Prognostic tests, 40, 223.
 Promotion, 43, 171, 259.

 Quality, 108.
 Quartile range, 327.

 Rate test, 63, 107.
 Recognition examination, 285.
 Regression equation, 348.
 Reliability, 201.
 Remedial instruction, 244, 251.
 Requirements of test construction, 64, 68 *ff.*

 Rice, J. M., 2, 3.
 Rugg, H. O., 321.

 Sampling, effect of, 330, 345.
 Scale, 15, 20, 106, 133.
 Scaled test, 62, 73, 78, 89, 92, 118.
 Secrist, Horace, 343.
 Skewness, 308.
 Spiral test, 63, 74.
 Standard deviation, 325.
 Standard units, 17, 21.
 Starch, Daniel, 28, 127.
 Stone, C. W., 6.
 Study tests, 51.
 Subjective, 26.

 Tests, definition, 20; selection of, 233.
 Testing by teacher, 237.
 Testing conditions, control of, 81 *ff.*
 Theisen, W. W., 196, 241.
 Thorndike, E. L., 2, 4, 24, 26, 122, 133, 186, 209, 216, 344.
 T-score, 151.
 Trabue, M. R., 138, 145.
 True score, 201.
 True-false examination, 283.

 Uniform tests, 62, 72, 113.

 Validity, 188, 194, 234.
 Validity of significance, 227.
 Van Wagenen, M. J., 125, 127, 241.
 Variable errors, 198, 243, 329, 344.

 Whipple, G. M., 321.
 Wood, Ben D., 284.
 Woodworth, R. S., 211.
 Woody, Clifford, 127, 145.

 Yule, G. Udny, 209, 349.

 Zero point, 101, 146, 150.
 Zirbes, Laura, 246.

RIVERSIDE TEXTBOOKS IN EDUCATION

General Educational Theory

PSYCHOLOGY FOR NORMAL SCHOOLS.

By L. A. AVERILL, Massachusetts State Normal School, Worcester.

EXPERIMENTAL EDUCATION.

By F. N. FREEMAN, University of Chicago.

HOW CHILDREN LEARN.

By F. N. FREEMAN.

THE PSYCHOLOGY OF THE COMMON BRANCHES.

By F. N. FREEMAN.

THE PRE-SCHOOL CHILD.

By ARNOLD GESELL, Ph.D., M.D., Director Yale Psycho-Clinic, Professor of Child Hygiene, Yale University.

DISCIPLINE AS A SCHOOL PROBLEM.

By A. C. PERRY, JR.

AN INTRODUCTION TO EDUCATIONAL SOCIOLOGY.

By W. R. SMITH, Kansas State Normal School.

TRAINING FOR EFFECTIVE STUDY.

By F. W. THOMAS, State Normal School, Fresno, California.

AN INTRODUCTION TO CHILD PSYCHOLOGY.

By C. W. WADDLE, Ph.D., Los Angeles State Normal School

History of Education

THE HISTORY OF EDUCATION.

By E. P. CUBBERLEY.

A BRIEF HISTORY OF EDUCATION.

By E. P. CUBBERLEY.

READINGS IN THE HISTORY OF EDUCATION.

By E. P. CUBBERLEY.

PUBLIC EDUCATION IN THE UNITED STATES.

By E. P. CUBBERLEY.

Administration and Supervision of Schools

HEALTHFUL SCHOOLS: HOW TO BUILD, EQUIP, AND MAINTAIN THEM.

By MAY AYRES, J. F. WILLIAMS, M.D., University of Cincinnati, and T. D. WOOD, A.M., M.D., Teachers College, Columbia University.

PUBLIC SCHOOL ADMINISTRATION.

By E. P. CUBBERLEY.

RURAL LIFE AND EDUCATION.

By E. P. CUBBERLEY.

A GUIDE TO EDUCATIONAL MEASUREMENTS.

By HARLAN C. HINES, Assistant Professor of Education, The University of Washington.

HEALTH WORK IN THE SCHOOLS.

By E. B. HOAG, M.D., and L. M. Terman, Leland Stanford Junior University.

INTRODUCTION TO THE THEORY OF EDUCATIONAL MEASUREMENTS.

By W. S. MONROE, University of Illinois.

MEASURING THE RESULTS OF TEACHING.

By W. S. MONROE.

EDUCATIONAL TESTS AND MEASUREMENTS.

By W. S. MONROE, J. C. DeVoss, Kansas State Normal School; and F. J. KELLY, University of Kansas.

THE SUPERVISION OF INSTRUCTION.

By H. W. NUTT, University of Kansas.

STATISTICAL METHODS APPLIED TO EDUCATION.

By H. O. RUGG, University of Chicago.

CLASSROOM ORGANIZATION AND CONTROL.

By J. B. SEARS, Leland Stanford Junior University.

A HANDBOOK FOR RURAL SCHOOL OFFICERS.

By N. D. SHOWALTER, Washington State Normal School.

THE HYGIENE OF THE SCHOOL CHILD.

By L. M. TERMAN.

THE MEASUREMENT OF INTELLIGENCE.

By L. M. TERMAN.

Test Material for the Measurement of Intelligence. Record Booklets for the Measurement of Intelligence.

THE INTELLIGENCE OF SCHOOL CHILDREN.

By L. M. TERMAN.

Methods of Teaching

TEACHING LITERATURE IN THE GRAMMAR GRADES AND HIGH SCHOOL.

By EMMA M. BOLENIUS.

HOW TO TEACH THE FUNDAMENTAL SUBJECTS.

By C. N. KENDALL and G. A. MIRICK.

HOW TO TEACH THE SPECIAL SUBJECTS.

By C. N. KENDALL and G. A. MIRICK.

SILENT AND ORAL READING.

By C. R. STONE.

THE TEACHING OF SCIENCE IN THE ELEMENTARY SCHOOL.

By G. H. TRAFTON, State Normal School, Mankato, Minnesota.

TEACHING IN RURAL SCHOOLS.

By T. J. WOOFER, University of Georgia.

Secondary Education

THE JUNIOR HIGH SCHOOL.

By THOS. H. BRIGGS, Columbia University.

THE TEACHING OF ENGLISH IN THE SECONDARY SCHOOL

By CHARLES SWAIN THOMAS.

PRINCIPLES OF SECONDARY EDUCATION.

By ALEXANDER INGLIS, Harvard University.

PROBLEMS OF SECONDARY EDUCATION.

By DAVID SNEDDEN, Columbia University.

HOUGHTON MIFFLIN COMPANY

1926 b

VOCATIONAL PREPARATION

THE VOCATIONAL GUIDANCE OF YOUTH

By Meyer Bloomfield

A monograph by the former Director of the Vocation Bureau of Boston.

YOUTH, SCHOOL, AND VOCATION

By Meyer Bloomfield

A first-hand presentation of the meaning and work of the vocational guidance movement.

CHOOSING A VOCATION

By Frank Parsons

This book is an indispensable manual for every vocational counselor.

THE PROBLEM OF VOCATIONAL EDUCATION

By David Snedden

The author is the Professor of Education, Teachers College, and one of the leaders in the movement for the closer adaptation of public schools to the actual needs of youth.

PREVOCATIONAL EDUCATION

By Frank M. Leavitt and Edith Brown

The first authoritative book to tell how the public schools may prepare pupils to select wisely the work to which they are best adapted.

THE PEOPLE'S SCHOOL

By Ruth Mary Weeks

A statement regarding the vocational training movement in this country and abroad.

VOCATIONS FOR GIRLS

By Mary A. Laselle and Katherine Wiley

Information as to conditions of work and the opportunities in the more common vocations open to girls with only a high-school education.

VOCATIONAL EDUCATION

By David Snedden, Ruth Mary Weeks, and Ellwood P. Cubberley

A combination of three volumes from the *Riverside Educational Monographs* treating different phases of vocational education,—theory, administration, and practice.

PRINCIPLES AND METHODS OF INDUSTRIAL EDUCATION

By William H. Dooley

This is a book for use in teacher training classes. There is an Introduction by Charles A. Prosser, and an equipment of thought stimulating questions, together with reading references and courses of study.

INDUSTRIAL EDUCATION: Its Problems, Methods, and Dangers

By Albert H. Leake

A study and criticism of the opportunities provided for the education of the industrial worker.

ESTABLISHING INDUSTRIAL SCHOOLS

By Harry Bradley Smith

A practical discussion of the steps to be taken in establishing industrial schools.

HOUGHTON MIFFLIN COMPANY

1908

DATE DUE

FEB. 10 1986

APR 08 '04

DEMCO 38-297

CINCINNATI BIBLE COLLEGE & SEM. LIBRARY
371.26 M753i main
Monroe, Walter Scot/An introduction to t



3 4320 00022 4693

371.26

#48376

he

al

THE CINCINNATI BIBLE SEMINARY LIBRARY

371.26 M753i

Ac. #48376

Monroe, Walter Scott

An introduction to the
theory of educational
measurements

